

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-143902
 (43)Date of publication of application : 28.05.1999

(51)Int.Cl.

G06F 17/30

(21)Application number : 09-309078

(71)Applicant : HITACHI LTD

(22)Date of filing : 11.11.1997

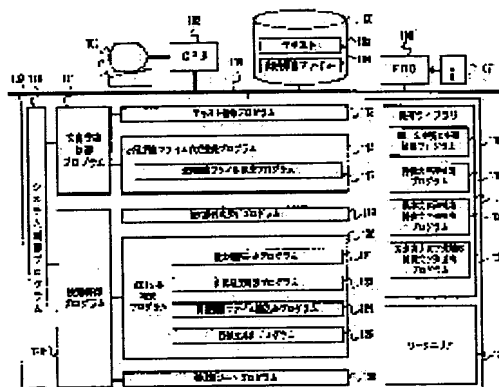
(72)Inventor : MATSUBAYASHI TADATAKA
 TADA KATSUMI
 OKAMOTO TAKUYA
 SUGAYA NATSUKO
 KAWASHITA YASUSHI

(54) SIMILAR DOCUMENT RETRIEVAL METHOD USING N-GRAM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a similar document retrieval system which can retrieve even for such languages as the Japanese have many character types at a high speed and with high accuracy.

SOLUTION: This system is provided with a step where the appearance frequency of a feature character string existing in a text 103 contained in a text data base is stored as an appearance frequency file 104, a step where the feature character string is extracted from the text that is designated by a user and a step where the appearance frequency of the feature character string is counted in the text designated by the user. Then, a similarity is calculated to the text designated by the user by using the appearance frequency stored in the file 104 and in the text designated by the user. Thus, the similar documents are retrieved by using the calculated similarity.



LEGAL STATUS

[Date of request for examination] 17.09.2002
 [Date of sending the examiner's decision of rejection]
 [Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
 [Date of final disposal for application]
 [Patent number]
 [Date of registration]
 [Number of appeal against examiner's decision of rejection]
 [Date of requesting appeal against examiner's decision of rejection]
 [Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平11-143902

(43)公開日 平成11年(1999) 5月28日

(51)Int.Cl.⁶

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/401

15/403

3 1 0 A

3 5 0 C

審査請求 未請求 請求項の数10 O L (全 27 頁)

(21)出願番号

特願平9-309078

(22)出願日

平成9年(1997)11月11日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 松林 忠孝

神奈川県川崎市幸区鹿島田890番地 株式
会社日立製作所情報・通信開発本部内

(72)発明者 多田 勝己

神奈川県川崎市幸区鹿島田890番地 株式
会社日立製作所情報・通信開発本部内

(72)発明者 岡本 卓哉

神奈川県川崎市幸区鹿島田890番地 株式
会社日立製作所情報・通信開発本部内

(74)代理人 弁理士 小川 勝男

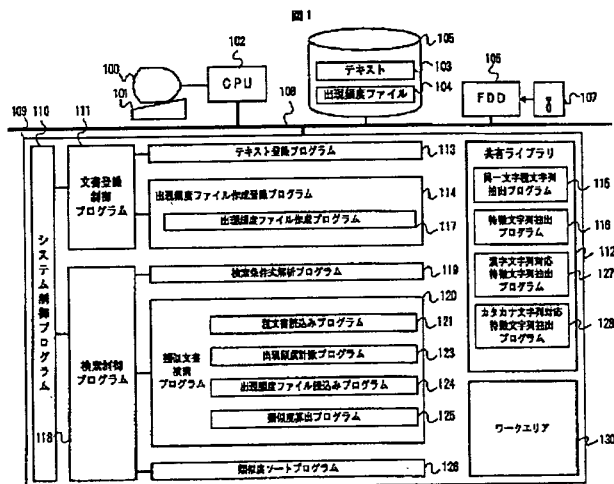
最終頁に続く

(54)【発明の名称】 n - g r a mを用いた類似文書検索方法

(57)【要約】

【課題】本発明の課題は、日本語のように文字種の多い言語に対しても、高速で高精度な類似文書検索システムを提供することである。

【解決手段】テキストデータベース中のテキスト103に存在する特徴文字列のそのテキスト103における出現頻度を出現頻度ファイル104として格納するステップと、ユーザが指定したテキストから特徴文字列を抽出するステップと、ユーザが指定したテキストにおける特徴文字列の出現頻度を計数するステップとを有し、出現頻度ファイル104とユーザが指定したテキストにおける出現頻度を用いてユーザが指定したテキストに対する類似度を算出し、算出された類似度を用いて文書を検索する。



【特許請求の範囲】

【請求項 1】文字情報をコードデータとして蓄積したテキストデータベースを対象に、ユーザが指定した文書と類似する文書を検索する類似文書検索方法において、ユーザが指定した文書のテキスト（指定テキストと呼ぶ）から所定の文字種の変わり目を境界として文字列を抽出する文字列抽出ステップと、予め定められた一つ以上の文字列の種類に応じて、その中から一つ以上の部分文字列を抽出する検索用部分文字列抽出ステップと、該指定テキストに対する該テキストデータベース中のテキストの類似度を所定の類似度算出式を用いて算出する類似度算出ステップを有することを特徴とした類似文書検索方法。

【請求項 2】請求項 1 記載の類似文書検索方法における前記文字列抽出ステップで、該指定テキストから抽出する文字列として、全ての文字種の変わり目を境界として同一文字種からなる文字列を抽出する同一文字種文字列抽出ステップを有することを特徴とした類似文書検索方法。

【請求項 3】請求項 2 記載の類似文書検索方法における前記検索用部分文字列抽出ステップで、全ての文字種に応じて予め定められた文字列長の部分文字列を検索用部分文字列として抽出する文字種別検索用部分文字列抽出ステップを有することを特徴とした類似文書検索方法。

【請求項 4】請求項 1、2 および 3 に記載の類似文書検索方法における前記検索用部分文字列抽出ステップで、予め定められた長さの文字列を検索用部分文字列として抽出するステップ、前記文字列抽出ステップで抽出された文字列そのものを検索用部分文字列として抽出するステップ、前記文字列抽出ステップで抽出された文字列とその部分文字列の指定テキストにおける出現頻度比を算出し、所定値を満たす部分文字列を検索用部分文字列として抽出するステップ、前記文字列抽出ステップで抽出された文字列から、予め作成しておいた、検索用部分文字列として抽出しない文字列を不要語として記載した排除文字列辞書に含まれない文字列を、検索用部分文字列として抽出するステップ、および、前記文字列抽出ステップで抽出された文字列から検索用部分文字列としては部分文字列を抽出しないステップ、のいずれか一つ、あるいは、それらを組み合わせることにより検索用部分文字列を抽出する検索用部分文字列抽出ステップを有することを特徴とした類似文書検索方法。

【請求項 5】請求項 1、2 および 3 に記載の類似文書検索方法における前記検索用部分文字列抽出ステップで、

予め定められた長さの文字列を検索用部分文字列として抽出する所定長文字列抽出ステップ、

前記文字列抽出ステップで抽出された文字列そのものを検索用部分文字列として抽出する最長文字列抽出ステップ、

前記文字列抽出ステップで抽出された文字列とその部分文字列の指定テキストにおける出現頻度比を算出し、所定値を満たす部分文字列を検索用部分文字列として抽出する高出現頻度比文字列抽出ステップ、

上記所定長文字列抽出ステップ、最長文字列抽出ステップおよび高出現頻度比文字列抽出ステップの中の少なくとも一つの抽出ステップで抽出された部分文字列から、予め作成しておいた、検索用部分文字列として抽出しない文字列を不要語として記載した排除文字列辞書に含まれる文字列を削除するステップ、

および、前記文字列抽出ステップで抽出された文字列から検索用部分文字列としては部分文字列を抽出しないステップ、

のいずれか一つ、あるいは、それらを組み合わせることにより検索用部分文字列を抽出する検索用部分文字列抽出ステップを有することを特徴とした類似文書検索方法。

【請求項 6】請求項 1、2、3 および 4 に記載の類似文書検索方法において、

前記検索用部分文字列抽出ステップで抽出された検索用部分文字列の重要度を、予め定められた算出式を用いて算出し、所定値を満たす検索用部分文字列を抽出する検索用部分文字列選択ステップを有することを特徴とした類似文書検索方法。

【請求項 7】請求項 5 記載の類似文書検索方法における前記検索用部分文字列選択ステップとして、前記検索用部分文字列抽出ステップにおいて抽出された検索用部分文字列の文字種類、文字列長、テキストデータベース内の出現文書数、指定テキストにおける出現頻度および該テキストにおける出現位置等の情報のいずれか一つ、あるいは、それらを組み合わせ、検索用部分文字列の重要度を算出する重要度算出ステップを有することを特徴とした類似文書検索方法。

【請求項 8】請求項 6 記載の類似文書検索方法において、

登録時に検索用部分文字列のテキストデータベース内の出現文書数を出現文書数ファイルとして保存する出現文書数ファイル作成ステップを有し、

検索時における前記重要度算出ステップにおいて、上記出現文書数ファイルから該検索用部分文字列の出現文書数を読み込む出現文書数ファイル読み込みステップを有することを特徴とした類似文書検索方法。

【請求項 9】請求項 6 および 7 に記載の類似文書検索方法において、

登録時に検索用部分文字列の重要度を予め定められた算

出式を用いて算出し、これを重要度ファイルとして保存する重要度ファイル作成ステップを有し、検索時における前記重要度算出ステップにおいて、上記重要度ファイルから該検索用部分文字列の重要度を読み込む重要度ファイル読み込みステップを有することを特徴とした類似文書検索方法。

【請求項 10】請求項 1～8 に記載の類似文書検索方法において、

登録時に検索用部分文字列のテキストデータベース内の各テキストにおける出現頻度を出現頻度ファイルとして保存する出現頻度ファイル作成ステップを有し、検索時における前記類似度算出ステップにおいて、上記出現頻度ファイルから出現頻度情報を読み込む出現頻度ファイル読み込みステップを有することを特徴とした類似文書検索方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ユーザが指定した文書と類似する文書を、文書データベースの中から検索する方法に関する。

【0002】

【従来の技術】近年、パーソナルコンピュータやインターネット等の普及に伴い、電子化文書が爆発的に増加しており、今後も加速度的に増大していくものと予想される。このような状況において、ユーザが所望する情報を含んだ文書を高速かつ効率的に検索したいという要求が高まってきている。

【0003】このような要求に応える技術として全文検索がある。全文検索では、検索対象文書をテキストとして計算機システムに登録してデータベース化し、この中からユーザが指定した検索文字列（以下、検索タームと呼ぶ）を含む文書を検索する。このように全文検索では、文書中の文字列そのものを対象として検索を行なうため、予めキーワードを付与し、このキーワードを手掛かりに検索する従来の検索システムとは異なり、検出漏れが原理的に生じないという特長がある。

【0004】しかし、ユーザが所望する情報を含んだ文書を的確に検索するためには、ユーザの検索意図を正確に表す複雑な検索条件式を入力する必要がある。

【0005】この複雑さを解消するために、ユーザが自分の所望する内容の文書（以下、種文書と呼ぶ）を指定し、その文書と類似する文書を検索する類似文書検索技術が注目されている。

【0006】類似文書検索の方法としては、例えば、「特開平 8 - 3 3 5 2 2 2 号公報」に、形態素解析等により種文書中に含まれる単語を抽出し、これを用いて類似文書を検索する技術（以下、従来技術 1 と呼ぶ）が開示されている。

【0007】また、「特開平 6 - 1 1 0 9 4 8 号公報」には、種文書中から連続する n 文字の文字列（以下、 n -g

ram と呼ぶ）を抽出し、これを用いて類似文書を検索する技術（以下、従来技術 2 と呼ぶ）が開示されている。

【0008】上記 2 つの従来技術について、その概要を説明する。

【0009】従来技術 1 では、形態素解析により種分書中に含まれる単語を抽出し、この単語を含む文書を類似文書として検索する。例えば、「この装置は地下水脈の観測にも使える。」という文書を種文書とする場合、形態素解析により単語辞書を参照して、「装置」「地下」「水脈」「観測」「使える」という単語が抽出される。この結果、「地下水脈を観測することによる地震の発生を予測する。」という文書を類似文書として検索することができる。しかし従来技術 1 では、単語の抽出に単語辞書を用いるため、次のような 2 つの問題が生じる。

【0010】まず、単語辞書に含まれていない単語は、種文書から検索用の単語として抽出されないためこの単語を含む文書を検索することができないという問題がある。このため、ユーザが所望する情報が新語で表され、これが単語辞書に含まれていない場合、目的の情報を含む文書を検索することができなくなる。

【0011】次に、ユーザが所望する情報を表す言葉が単語辞書に含まれる場合でも、単語の抽出の仕方によっては検索漏れが生じてしまう。例えば、上記の「この装置は地下水脈の観測にも使える。」という種文書からは、「装置」「地下」「水脈」「観測」「使える」という単語が抽出される。しかし、「地下水」という単語が抽出されないため「地下水の大量汲み上げで地盤沈下地域が拡大した。」という文書は、類似文書として検索することができないという問題がある。

【0012】以上が従来技術 1 の問題点である。

【0013】この問題を解決するために、上記従来技術 2 が提案されている。これは、 n -gram を用いた類似文書検索方法である。

【0014】以下、文書 1 「新開発の心電計による発作時の心電図」、文書 2 「新しいソフトウェアの開発作業」、および文書 3 「ソフト開発を支援するソフトウェア」が登録されているデータベースを対象に、 n -gram の n の値を 2 として、ユーザが文書 2 を種文書に指定した場合を例に、従来技術 2 の具体的な処理方法を説明し、その問題点を述べる。

【0015】まず、データベース中の文書 1～文書 3 から 2-gram を抽出する。

【0016】

【表 1】

表1

No.	重複排除2-gram	出現頻度	ウェイト
1-1	新開	1	(1/16)= 0.063
1-2	開発	1	
1-3	発の	1	
1-4	の心	2	(2/16)= 0.125
1-5	心電	2	
1-6	電計	1	
1-7	計に	1	0.063
1-8	によ	1	
1-9	る発	1	
1-10	る発	1	0.063
1-11	発作	1	
1-12	作時	1	
1-13	時の	1	0.063
1-14	電図	1	
(総出現頻度)		16	

【0017】表1に、文書1に出現する2-gramの中から重複を排除して抽出した2-gram（以下、重複排除2-gramと呼ぶ）を示す。次に、これらの2-gramの各々に対しウェイトを計算する。このウェイトは各々の2-gramの出現頻度をその文書中に出現する2-gramの重複を含めた総出現頻度で割ることによって求める。ウェイトは各々の2-gramのその文書内における出現割合を表し、この値が大きい2-gramほどその文書に頻繁に出現することを意味する。文書2および文書3に対しても同様の処理を施し、それぞれウェイトを求める。表2および表3にこの処理結果を示す。

【0018】

【表2】

表2

No.	重複排除2-gram	出現頻度	ウェイト
2-1	新し	1	(1/13)= 0.077
2-2	しい	1	
2-3	いソ	1	
2-4	ソフ	1	0.077
2-5	フト	1	
2-6	トウ	1	
2-7	ウェ	1	0.077
2-8	エア	1	
2-9	アの	1	
2-10	の開	1	0.077
2-11	開発	1	
2-12	発作	1	
2-13	作業	1	0.077
(総出現頻度)		13	

【0019】

【表3】

表3

No.	重複排除2-gram	出現頻度	ウェイト
3-1	ソフ	2	(1/15)= 0.133
3-2	フト	2	
3-3	ト開	1	(1/13)= 0.067
3-4	開発	1	
3-5	発を	1	
3-6	を支	1	0.067
3-7	支援	1	
3-8	握す	1	
3-9	する	1	0.067
3-10	るソ	1	
3-11	トウ	1	
3-12	ウェ	1	0.067
3-13	エア	1	
(総出現頻度)		15	

【0020】その後、データベース中の文書間の共通性を除去する。ここでは、まず、データベース中に存在する2-gramの中で重複を排除した2-gramについて、その共通性ウェイトを算出する。この共通性ウェイトは、データベース中の全文書に関する各2-gramのウェイトの総和を、全文書数で割ることによって算出する。共通性ウェイトは、各2-gramのデータベース全体における出現割合を表し、この値が大きい2-gramほどデータベース中のどの文書にも共通して出現することを意味する。

【0021】

【表4】

表4

No.	重複排除2-gram	ウェイト			共通性ウェイト
		文 1	文 2	文書3	
4-1	新開	0.063	0.077	0.067	(0.063+0.0+0.0)/3= 0.021
4-2	開発	0.063			0.069
4-3	発の	0.063			0.021
4-4	の心	0.125			0.042
4-5	心電	0.125			0.042
4-6	電計	0.063			0.021
4-7	計に	0.063			0.021
4-8	によ	0.063			0.021
4-9	よる	0.063			0.021
4-10	る発	0.063			0.021
4-11	発作	0.063	0.077		0.046
4-12	作時	0.063			0.021
4-13	時の	0.063			0.021
4-14	電図	0.063			0.021
4-15	新しい	0.077	0.077		0.026
4-16	しいソ	0.077			0.026
4-17	ソフト	0.077			0.026
4-18	ソフト	0.077			0.133
4-19	フトウ	0.077			0.133
4-20	ウェア	0.077			0.067
4-21	エア	0.077			0.067
4-22	アの	0.077			0.067
4-23	の開	0.077			0.067
4-24	作業	0.077			0.067
4-25	ト開	0.077			0.067
4-26	発を	0.077			0.067
4-27	を支	0.077			0.067
4-28	支援	0.077			0.067
4-29	握す	0.077			0.067
4-30	するソ	0.077			0.067
4-31					0.022
4-32					0.022

【0022】表4に、文書1、文書2および文書3の間の共通性ウェイトを示す。

【0023】例えば、2-gram「新開」の共通性ウェイトは、 $(0.063+0.0+0.0)/3=0.021$ である。ここで2-gram「新開」は文書2および文書3に出現していないのでウェイトはそれぞれ「0.0」となっている。2-gram「開発」の共通性ウェイトは、 $(0.063+0.077+0.067)/3=0.069$ である。

【0024】上述したように、共通性ウェイトは各n-gramのウェイトの平均値である。

【0025】この共通性ウェイトを各n-gramのウェイトから差し引くことにより、データベース中の文書間の共通性を除去する（この値を従来技術2では、正規化ウェイトと呼んでいる）。正規化ウェイトは、データベース

における各n-gramの出現偏りを表し、この値が大きいn-gramほどある特定の文書に偏って出現することを意味する。

【0026】もし、あるn-gramが全ての文書に同じ割合で出現していれば、ウェイトと共通性ウェイトは同じ値となるため、正規化ウェイトは「0」となる。つまり、どの文書においても同じような割合で出現するn-gramに関しては、ウェイトが限りなく「0」に近づくことになる。

【0027】表5、表6および表7に、文書1、文書2および文書3の正規化ウェイトを示す。

【0028】

【表5】

表5

No.	重複排除2-gram	ウェイト	共通性ウェイト	正規化ウェイト
5-1	新開	0.063	0.021	$(0.063-0.021)=$ 0.042
5-2	開発	0.063	0.069	-0.006
5-3	発の	0.063	0.042	0.021
5-4	の心	0.125	0.042	0.083
5-5	心電	0.125	0.021	0.104
5-6	電計	0.063	0.021	0.042
5-7	計に	0.063	0.021	0.042
5-8	によ	0.063	0.021	0.042
5-9	よる	0.063	0.021	0.042
5-10	る発	0.063	0.021	0.042
5-11	発作	0.063	0.046	0.016
5-12	作時	0.063	0.021	0.042
5-13	時の	0.063	0.021	0.042
5-14	電図	0.063	0.021	0.042

【0029】

【表6】

表6

No.	重複排除 2-gram	ウェイト	共通性 ウェイト	正規化ウェイト
6-1	新し	0.077	0.026	(0.063-0.026)= 0.051
6-2	しい	0.077	0.026	
6-3	いソ	0.077	0.026	
6-4	ソフ	0.077	0.070	
6-5	フト	0.077	0.070	
6-6	トウ	0.077	0.048	
6-7	ウェ	0.077	0.048	
6-8	エア	0.077	0.048	
6-9	アの	0.077	0.026	
6-10	の開	0.077	0.026	
6-11	開発	0.077	0.069	
6-12	発作	0.077	0.046	
6-13	作業	0.077	0.026	

【0030】

【表7】

表7

No.	重複排除 2-gram	ウェイト	共通性 ウェイト	正規化ウェイト
7-1	ソフ	0.133	0.070	(0.133-0.070)= 0.063
7-2	フト	0.133	0.070	
7-3	ト開	0.067	0.022	
7-4	開発	0.067	0.069	
7-5	発を	0.067	0.022	
7-6	を支	0.067	0.022	
7-7	支援	0.067	0.022	
7-8	振す	0.067	0.022	
7-9	する	0.067	0.022	
7-10	るソ	0.067	0.022	
7-11	トウ	0.067	0.048	
7-12	ウェ	0.067	0.048	
7-13	エア	0.067	0.048	

【0031】 以上のようにして得られた正規化ウェイトを用いて、ユーザが種文書として指定した文書とデータベース中の全文書との類似性を求め、これを類似度として表わす。文書番号を*i*とすると、文書*i*の類似度*S(i)*は、以下に示す式(1)によって求められる。

【0032】

【数1】

数式1

$$S(i) = \frac{\sum_{j=1}^n (U(j) \times R(j))}{\sqrt{\sum_{j=1}^n U(j)^2 \times \sum_{j=1}^n R(j)^2}} \quad \dots \quad 1$$

【0033】 ここで、*U(j)*は種文書中の*j*番目の*n*-gramの正規化ウェイトを示し、*R(j)*はデータベース中文書の*j*番目の*n*-gramの正規化ウェイトを示す。また、*n*はデータベース中の全文書数を表わす。この式を用いてデータベース中の全ての文書の類似度を算出すると以下のようになる。

【0034】

S(1) = 0.018*S*(2) = 1.0*S*(3) = 0.119

最後に、得られた類似度の降順に文書を出力する。この例では、文書2、文書3、文書1の順で出力されること

になる。

【0035】 以上が、従来技術2の具体的な処理内容である。このように従来技術2では、単語辞書に基づく形態素解析を用いることなく種文書に類似する文書を検索することができるため、従来技術1における2つの問題点を解決することができる。

【0036】 しかし、この従来技術2には次のような2つの問題点がある。

【0037】 まず、第一の問題点は、種文書から抽出される*n*-gram数が膨大になるため、検索に長大な時間を要してしまうという問題である。例えば、1,000文字からなる種文書から全ての2-gramを抽出した場合、999個の2-gramが抽出されることになる。そのため、抽出した全ての2-gramを類似検索に用いる従来技術2の方法では、1個の2-gramの検索が0.1秒で済んだとしても、999個の2-gramでは99.9秒、すなわち約1分40秒も検索時間が掛かってしまうことになる。

【0038】 また、第二の問題点は、全ての*n*-gramを用いて類似文書を検索するため、検索結果にノイズが含まれるという問題である。

【0039】 以下、この問題点を、文書1～文書3が登録されている前記データベースに、文書4「新しいソフトクリーム券の配布作業」を追加した場合を例に、具体的に説明する。

【0040】 本例では、文書2が、種文書としてユーザ

に指定されたものとする。

【0041】まず、文書4から2-gramを抽出し、ウェイトを求めた結果を表8に示す。

【0042】

【表8】

表8

No.	重複排除2-gram	出現頻度	ウェイト
8-1	新し	1	(1/15)= 0.067
8-2	しい	1	0.067
8-3	いソ	1	0.067
8-4	ソフ	1	0.067
8-5	フト	1	0.067
8-6	トク	1	0.067
8-7	クリ	1	0.067
8-8	リー	1	0.067
8-9	ーム	1	0.067
8-10	ム券	1	0.067
8-11	券の	1	0.067
8-12	の配	1	0.067
8-13	配付	1	0.067
8-14	付作	1	0.067
8-15	作業	1	0.067
(総出現頻度)		15	

【0043】この文書4のウェイトと表1～表3に示した文書1～文書3のウェイトを用いて、共通性ウェイトを算出する。

【0044】

【表9】

表9

No.	重複排除2-gram	ウェイト				共通性ウェイト
		文書1	文書2	文書3	文書4	
9-1	新開	0.063				0.016
9-2	開発	0.063	0.077	0.067		0.052
9-3	発の	0.063				0.016
9-4	の心	0.125				0.031
9-5	心電	0.125				0.031
9-6	電計	0.063				0.016
9-7	計に	0.063				0.016
9-8	によ	0.063				0.016
9-9	よる	0.063				0.016
9-10	る発	0.063				0.016
9-11	発作	0.063	0.077			0.035
9-12	作時	0.063				0.016
9-13	時の	0.063				0.016
9-14	電図	0.063				0.016
9-15	新し		0.077		0.067	0.036
9-16	しい		0.077		0.067	0.036
9-17	いソ		0.077		0.067	0.036
9-18	ソフ		0.077	0.133	0.067	0.069
9-19	フト		0.077	0.133	0.067	0.069
9-20	トウ		0.077	0.067		0.036
9-21	ウェ		0.077	0.067		0.036
9-22	エア		0.077	0.067		0.036
9-23	アの		0.077			0.019
9-24	の開		0.077			0.019
9-25	作業		0.077		0.067	0.036
9-26	ト開			0.067		0.017
9-27	発を			0.067		0.017
9-28	を支			0.067		0.017
9-29	支援			0.067		0.017
9-30	援す			0.067		0.017
9-31	する			0.067		0.017
9-32	るソ			0.067		0.017
9-33	トク				0.067	0.017
9-34	クリ				0.067	0.017
9-35	リー				0.067	0.017
9-36	ーム				0.067	0.017
9-37	ム券				0.067	0.017
9-38	券の				0.067	0.017
9-39	の配				0.067	0.017
9-40	配布				0.067	0.017
9-41	布作				0.067	0.017

【0045】表9に、文書1～文書4の間の共通性ウェイトを示す。例えば、2-gram「開発」の共通性ウェイトは、 $(0.063+0.077+0.067+0.000)/4=0.052$ となる。次

に、この共通性ウェイトを各文書の重複排除2-gramのウェイトから差し引くことにより、データベース中の文書間の共通性を除去した正規化ウェイトを求める。

【0046】

【表10】

表10

No.	重複排除 2-gram	ウェイト	共通性 ウェイト	正規化ウェイト	
10-1	新開	0.063	0.016	(0.063-0.016)= -0.006	0.047
10-2	開発	0.063	0.068		0.047
10-3	発の	0.063	0.016		0.094
10-4	の心	0.125	0.031		0.094
10-5	心電	0.125	0.031		0.047
10-6	電計	0.063	0.016		0.047
10-7	計に	0.063	0.016		0.047
10-8	によ	0.063	0.016		0.047
10-9	よる	0.063	0.016		0.047
10-10	る発	0.063	0.016		0.047
10-11	発作	0.063	0.035		0.028
10-12	作時	0.063	0.016		0.047
10-13	時の	0.063	0.016		0.047
10-14	電図	0.063	0.016		0.047

【0047】

【表11】

表11

No.	重複排除 2-gram	ウェイト	共通性 ウェイト	正規化ウェイト	
11-1	新し	0.077	0.036	(0.077-0.036)= 0.041	0.041
11-2	しい	0.077	0.036		0.041
11-3	いソ	0.077	0.036		0.041
11-4	ソフ	0.077	0.069		0.008
11-5	フト	0.077	0.069		0.008
11-6	トウ	0.077	0.036		0.041
11-7	ウェ	0.077	0.036		0.041
11-8	エア	0.077	0.036		0.041
11-9	アの	0.077	0.019		0.058
11-10	の開	0.077	0.036		0.041
11-11	開発	0.077	0.068		0.009
11-12	発作	0.077	0.035		0.042
11-13	作業	0.077	0.036		0.041

【0048】

【表12】

表12

No.	重複排除 2-gram	ウェイト	共通性 ウェイト	正規化ウェイト
12-1	ソフ	0.133	0.069	(0.133-0.069)= 0.064
12-2	フト	0.133	0.069	
12-3	ト開	0.067	0.017	
12-4	開発	0.067	0.068	
12-5	発を	0.067	0.017	
12-6	を支	0.067	0.017	
12-7	支援	0.067	0.017	
12-8	援す	0.067	0.017	
12-9	する	0.067	0.017	
12-10	るソ	0.067	0.017	
12-11	トウ	0.067	0.036	
12-12	ウェ	0.067	0.036	
12-13	エア	0.067	0.036	

【0049】

【表13】

表 13

No.	重複排除 2-gram	ウェイト	共通性 ウェイト	正規化ウェイト
13-1	新し	0.067	0.036	(0.067-0.036)= 0.031
13-2	しい	0.067	0.036	0.031
13-3	いソ	0.067	0.036	0.031
13-4	ソフ	0.067	0.069	-0.003
13-5	フト	0.067	0.069	-0.003
13-6	トク	0.067	0.017	0.050
13-7	クリ	0.067	0.017	0.050
13-8	リー	0.067	0.017	0.050
13-9	ーム	0.067	0.017	0.050
13-10	ム券	0.067	0.017	0.050
13-11	券の	0.067	0.017	0.050
13-12	の配	0.067	0.017	0.050
13-13	配布	0.067	0.017	0.050
13-14	布作	0.067	0.017	0.050
13-15	作業	0.067	0.036	0.031

【0050】表10～表13に文書1～文書4における2-gramの正規化ウェイトを示す。これらを用いて、種文書である文書2に対する各文書の類似度を式(1)を用いて算出すると、

$$S(1) = 0.036$$

$$S(2) = 1.0$$

$$S(3) = 0.179$$

$$S(4) = 0.190$$

となる。

【0051】ここで、文書3は文書2と同様にソフトウェアに関する文書であるにも関わらず、関係のない文書4の方が文書2に類似していると判断されてしまっている。これは、文書2の「ソフトウェア」から抽出される「ソフ」「フト」が、文書4の「ソフトクリーム」からも抽出されることによる。n-gramは単語のように意味的にまとまった単位の文字列ではないため、同じn-gramであっても文書内で同じ意味を表現しているとは限らない。そのため、この例のように全く関係のない文書が高い類似度を持つ文書として探し出されてしまうという問題がある。

【0052】

【発明が解決しようとする課題】こうした従来技術の問題に対し、本発明では以下の課題を解決することを目的とする。

【0053】(1) 検索精度の高い類似文書検索方法を提供する。

【0054】(2) 日本語のように文字種の多い言語に対しても、高速に類似文書検索が行える方法を提供する。

【0055】

【課題を解決するための手段】上記課題を解決するために、本発明による文書検索方法では、以下に示すステップで種文書と類似する文書を検索する。

【0056】すなわち、本発明による文書検索方法では、文書の登録処理として、(ステップ1)登録対象文書を読み込む文書読み込みステップ、(ステップ2)上記文書読み込みステップで読み込んだ登録対象文書の文字列

を、漢字やカタカナ等の文字種境界で分割し、同一文字種で構成される文字列(以下、同一文字種文字列と呼ぶ)として抽出する同一文字種文字列抽出ステップ、

(ステップ3)上記同一文字種文字列抽出ステップで抽出した同一文字種文字列に対して、その文字種を判定し、漢字ならば予め定められた長さの文字列を自立語の可能性のあるもの(以下、特徴文字列と呼ぶ)として、そこから抽出し、カタカナや英字ならば同一文字種文字列そのものを特徴文字列として抽出し、それ以外の文字種ならば特徴文字列としては抽出を行わない登録用特徴文字列抽出ステップ、(ステップ4)上記登録用特徴文字列抽出ステップで抽出した特徴文字列に関して、登録対象文書内における出現頻度を計数する出現頻度計数ステップ、(ステップ5)上記出現頻度計数ステップで計数した出現頻度を該当する出現頻度ファイルに格納する出現頻度ファイル作成登録ステップ、を有し、種文書に類似する文書の検索処理として、(ステップ6)種文書を読み込む文書読み込みステップ、(ステップ7)上記種文書読み込みステップにおいて読み込んだ種文書の文字列を文字種境界で分割し、同一文字種文字列として抽出する同一文字種文字列抽出ステップ、(ステップ8)上記同一文字種文字列抽出ステップで抽出した同一文字種文字列に対して、その文字種を判定し、漢字ならば予め定められた長さの文字列を特徴文字列としてそこから抽出し、カタカナや英字ならば同一文字種文字列そのものを特徴文字列として抽出し、それ以外の文字種ならば特徴文字列としては抽出を行わない検索用特徴文字列抽出ステップ、(ステップ9)上記検索用特徴文字列抽出ステップで抽出した特徴文字列に関して、種文書内の出現頻度を計数する出現頻度計数ステップ、(ステップ10)上記出現頻度計数ステップで抽出した全ての特徴文字列に対して、前記出現頻度ファイルを読み込み、データベース内の各文書における出現頻度を取得する出現頻度取得ステップ、(ステップ11)上記出現頻度取得ステップで抽出した特徴文字列に関し、上記出現頻度計数ステップで計数した種文書内の出現頻度と、上記出現頻度取得ステップで取得したデータベース内の各文書にお

ける出現頻度を用いて、予め定められた算出式に基づいて種文書とデータベース内の各文書との類似度を算出する類似度算出ステップ、(ステップ12)上記類似度算出ステップで算出した類似度の降順に、文書の一覧を表示する検索結果表示ステップを有する。

【0057】上記文書検索方法を用いた本発明の原理を、以下に説明する。

【0058】文書を登録する際には、(ステップ1)～(ステップ5)を実行する。まず、(ステップ1)で登録対象となる文書を読み込む。次に、(ステップ2)において、(ステップ1)で読み込んだ登録対象文書中の文字列を、漢字やカタカナ等の文字種境界で分割し、同一文字種からなる文字列を抽出する。例えば、前記の文書4「新しいソフトクリーム券の配布作業」という文書からは、「新」「しい」「ソフトクリーム」「券」「の」「配布作業」という6個の同一文字種文字列が抽出される。

【0059】次に、(ステップ3)において、(ステップ2)で抽出した同一文字種文字列について、その文字種を判定し、漢字ならば予め定められた長さの文字列を特徴文字列としてそこから抽出し、カタカナや英字ならば同一文字種文字列そのものを特徴文字列として抽出し、それ以外の文字種ならば特徴文字列としては抽出を行わない。例えば、予め漢字文字列から2-gramを抽出するものと定められている場合には、上記(ステップ2)における同一文字種文字列からは、「ソフトクリーム」「配布」「布作」「作業」が特徴文字列として抽出される。

【0060】次に、(ステップ4)において、(ステップ3)で抽出した特徴文字列の登録対象文書内における出現頻度を計数する。例えば、上記の文書4「新しいソフトクリーム券の配布作業」という文書では、特徴文字列「ソフトクリーム」が1回出現し、「作業」は1回出現するという情報が得られる。

【0061】次に、(ステップ5)において、先に(ステップ4)で計数した特徴文字列の出現頻度を該当する出現頻度ファイルに格納する。図2に出現頻度ファイルの例を示す。本図に示した出現頻度ファイルは、表1、表2、表3および表8に示した文書1～文書4を登録した場合の例である。

【0062】検索時には、(ステップ6)～(ステップ12)からなる類似文書検索ステップを実行する。

【0063】まず、(ステップ6)において、種文書として文書2を読み込む。

【0064】次に、(ステップ7)において、(ステップ6)で読み込んだ種文書(文書2)の文字列を文字種境界で分割し、同一文字種文字列を抽出する。

【0065】次に、(ステップ8)において、上記(ステップ7)で抽出した同一文字種文字列から、登録時の(ステップ3)と同様の方法で特徴文字列を抽出する。

図3に文書2が種文書として指定された場合の(ステップ8)の特徴文字列抽出処理の概要を示す。本図では、同一文字種文字列が漢字の場合には、2-gramを抽出するものとしている。文書2から全ての2-gramを抽出した場合には、13種類の2-gramが抽出されていたのに対し、本方法では、「ソフトウェア」「開発」「発作」「作業」の4種類の特徴文字列に削減することができている。このように、全てのn-gramを抽出する前述した従来技術2に比べ、本発明では抽出する特徴文字列の種類を大幅に削除できることになる。

【0066】次に、(ステップ9)において、(ステップ8)で抽出した特徴文字列の種文書内における出現頻度を計数する。そして、(ステップ10)において、(ステップ8)で抽出した特徴文字列に関して、前述した出現頻度ファイルを参照し、データベース内の各文書における出現頻度を得る。そして、(ステップ11)において、(ステップ8)で抽出した特徴文字列に対して、(ステップ9)と(ステップ10)で計数した種文書内における出現頻度と、データベース内の各文書における出現頻度を基に、類似度を算出する。類似度の算出式には、従来技術2で示した式(1)を用いてもよい。式(1)を用いて、文書2が種文書として指定された場合の類似度を算出すると、次のようになる。

【0067】

$S(1)=0.077$

$S(2)=1.0$

$S(3)=0.263$

$S(4)=0.148$

この結果、(ステップ12)で、文書を類似度の降順に表示すると、文書2、文書3、文書4および文書1の順に表示される。この類似度算出結果($S(1)=0.077$ 、 $S(2)=1.0$ 、 $S(3)=0.263$ 、 $S(4)=0.148$)は、従来技術2による類似度算出結果($S(1)=0.036$ 、 $S(2)=1.0$ 、 $S(3)=0.179$ 、 $S(4)=0.190$)とは異なり、文書2に類似した順に、類似度が正しく算出されることになる。

【0068】以上のように、本発明の類似文書検索方法によれば、分かち書きのない日本語のような文書に対して、類似文書検索を行なっても、従来技術1のような単語辞書を用いることなく種文書から文字列を機械的に抽出するため、従来技術2のようにどんな単語についても漏れのない検索を行なうことが可能となる。また、従来技術2のように文書中から単純にn-gramを抽出するのではなく、文字種に応じて特徴文字列を抽出することにより、意味のまとまった文字列を用いて検索を行なうことができるため、高精度な類似文書検索を実現することができる。さらに、全n-gramを抽出する従来技術2に比べ、抽出する文字列の種類が大幅に削減されるため、高速に類似文書を検索することができるようになる。

【0069】

【発明の実施の形態】以下、本発明の第一の実施例について図1を用いて説明する。

【0070】本発明を適用した類似文書検索システムの第一例は、ディスプレイ100、キーボード101、中央演算処理装置(CPU)102、磁気ディスク装置105、フロッピディスクドライブ(FDD)106、主メモリ109およびこれらを結ぶバス108から構成される。

【0071】磁気ディスク装置105は二次記憶装置の一つであり、テキスト103、出現頻度ファイル104が格納される。FDD106を介してフロッピディスク107に格納されている情報が、主メモリ109あるいは磁気ディスク装置105へ読み込まれる。

【0072】主メモリ109には、システム制御プログラム110、文書登録制御プログラム111、共有ライブラリ112、テキスト登録プログラム113、出現頻度ファイル作成登録プログラム114、検索制御プログラム118、検索条件式解析プログラム119、類似文書検索プログラム120および類似度ソートプログラム126が格納されるとともにワークエリア130が確保される。

【0073】共有ライブラリ112は、同一文字種文字列抽出プログラム115、特徴文字列抽出プログラム116、漢字文字列対応特徴文字列抽出プログラム127およびカタカナ文字列対応特徴文字列抽出プログラム128で構成される。

【0074】出現頻度ファイル作成登録プログラム114は、出現頻度ファイル作成プログラム117で構成されると共に、後述するように同一文字種文字列抽出プログラム115と特徴文字列抽出プログラム116を呼び出す構成をとる。

【0075】類似文書検索プログラム120は、種文書読込みプログラム121、同一文字種文字列抽出プログラム115、出現頻度計数プログラム123、出現頻度ファイル読込みプログラム124および類似度算出プログラム125で構成されると共に、後述するように特徴文字列抽出プログラム116を呼び出す構成をとる。

【0076】文書登録制御プログラム111および検索制御プログラム118は、ユーザによるキーボード101からの指示に応じてシステム制御プログラム110によって起動され、それぞれテキスト登録プログラム113および出現頻度ファイル作成登録プログラム114の制御と、検索条件式解析プログラム119、類似文書検索プログラム120および類似度ソートプログラム126の制御を行なう。

【0077】以下、本実施例における類似文書検索システムの処理手順について説明する。

【0078】まず、システム制御プログラム110の処理手順について図4のPAD(Problem Analysis Diagram)図を用いて説明する。

【0079】システム制御プログラム110は、まずステップ400で、キーボード101から入力されたコマンドを解析する。

【0080】そしてステップ401で、この結果が登録実行のコマンドであると解析された場合には、ステップ402で文書登録制御プログラム111を起動して、文書の登録を行なう。

【0081】またステップ403で、検索実行のコマンドであると解析された場合には、ステップ404で検索制御プログラム118を起動して、類似文書の検索を行なう。

【0082】以上が、システム制御プログラム110の処理手順である。

【0083】次に、図4に示したステップ402でシステム制御プログラム110により起動される文書登録制御プログラム111の処理手順について、図5のPAD図を用いて説明する。

【0084】文書登録制御プログラム111は、まずステップ500でテキスト登録プログラム113を起動し、FDD106に挿入されたフロッピディスク107から登録すべき文書のテキストデータをワークエリア130に読み込み、これをテキスト103として磁気ディスク装置105に格納する。テキストデータは、フロッピディスク107を用いて入力するだけに限らず、通信回線やCD-ROM装置(図1には示していない)等を用いて他の装置から入力するような構成を取ることも可能である。

【0085】次に、ステップ501で出現頻度ファイル作成登録プログラム114を起動し、磁気ディスク装置105に格納されているテキスト103を読み出し、その中の各文書における出現頻度ファイル104を作成し、磁気ディスク装置105に格納する。

【0086】以上が、文書登録制御プログラム111の処理手順である。

【0087】次に、図5に示したステップ501で文書登録制御プログラム111により起動される出現頻度ファイル作成登録プログラム114の処理手順について、図6のPAD図を用いて説明する。

【0088】出現頻度ファイル作成登録プログラム114は、まずステップ600で同一文字種文字列抽出プログラム115を起動し、テキスト103をワークエリア130に読み込み、文字種境界でその文字列を分割することにより同一文字種文字列を抽出し、ワークエリア130に格納する。

【0089】次に、ステップ601において、特徴文字列抽出プログラム116を起動し、ワークエリア130に格納されている同一文字種文字列から特徴文字列を抽出し、同じくワークエリア130に格納する。

【0090】そして、ステップ602において、出現頻度ファイル作成プログラム117を起動し、ワークエリ

ア130に格納されている特徴文字列を参照して、その出現頻度を計数し、出現頻度ファイル104として磁気ディスク装置105に格納する。

【0091】以上が、出現頻度ファイル作成登録プログラム114の処理手順である。

【0092】次に、図6に示したステップ601において出現頻度ファイル作成登録プログラム114により起動される特徴文字列抽出プログラム116の処理手順について、図7のPAD図を用いて説明する。

【0093】特徴文字列抽出プログラム116は、同一文字種文字列抽出プログラム115により抽出された同一文字種文字列の数を調べ、全ての同一文字種文字列についてステップ701以降を繰り返し実行する（ステップ700）。

【0094】ステップ701では、ワークエリア130に格納されている同一文字種文字列の文字種を判定し、その文字種が漢字の場合にはステップ702を実行し、カタカナの場合には、ステップ703を実行する。

【0095】ステップ702では、後述する漢字文字列対応特徴文字列抽出プログラム127を起動し、漢字文字列から特徴文字列を抽出する。

【0096】ステップ703では、同様に後述するカタカナ文字列対応特徴文字列抽出プログラム128を起動し、カタカナ文字列から特徴文字列を抽出する。

【0097】以上が、特徴文字列抽出プログラム116の処理手順である。

【0098】次に、図7に示したステップ702で特徴文字列抽出プログラム116により起動される漢字文字列対応特徴文字列抽出プログラム127の処理手順について、図8のPAD図を用いて説明する。

【0099】漢字文字列対応特徴文字列抽出プログラム127では、ステップ800において、同一文字種文字列抽出プログラム115により抽出されワークエリア130に格納されている漢字文字列を取得する。そして、ステップ801において、上記ステップ800で取得した漢字文字列の先頭から一文字ずつずらしながら、n-gram（nの値は、予め定めておく）を特徴文字列として抽出する。

【0100】以上が、漢字文字列対応特徴文字列抽出プログラム127の処理手順である。

【0101】次に、図7に示したステップ703で特徴文字列抽出プログラム116により起動されるカタカナ文字列対応特徴文字列抽出プログラム128の処理手順について、図9のPAD図を用いて説明する。

【0102】カタカナ文字列対応特徴文字列抽出プログラム128では、ステップ900において、同一文字種文字列抽出プログラム115により抽出されワークエリア130に格納されているカタカナ文字列を取得する。

【0103】そして、ステップ901において、上記ステップ900で取得したカタカナ文字列そのものを特徴

文字列として抽出する。

【0104】以上が、カタカナ文字列対応特徴文字列抽出プログラム128の処理手順である。

【0105】以下に、図7に示した特徴文字列抽出プログラム116の処理手順について具体例を用いて説明する。

【0106】まず、図7の特徴文字列抽出プログラム116のステップ702における漢字文字列対応特徴文字列抽出プログラム127と、ステップ703におけるカタカナ文字列対応特徴文字列抽出プログラム128の処理手順について、図10～図12の例を用いて説明する。漢字文字列対応特徴文字列抽出プログラム127とカタカナ文字列対応特徴文字列抽出プログラム128は特徴文字列抽出プログラム116によって起動される。このとき、同一文字種文字列抽出プログラム115によって抽出された同一文字種文字列が漢字文字列対応特徴文字列抽出プログラム127とカタカナ文字列対応特徴文字列抽出プログラム128へワークエリア130を介して渡される。

【0107】図10は文書1、文書2、文書3および文書4からなるテキスト103から、同一文字種文字列抽出プログラム115により同一文字種文字列が抽出された結果を示したものである。例えば、文書2「新しいソフトウェアの開発作業」からは「新」「しい」「ソフトウェア」「の」「開発作業」という5個の同一文字種文字列が抽出される。

【0108】この抽出された同一文字種文字列の文字種にしたがって、特徴文字列抽出プログラム116は、漢字文字列対応特徴文字列抽出プログラム127あるいはカタカナ文字列対応特徴文字列抽出プログラム128を起動する。

【0109】漢字文字列対応特徴文字列抽出プログラム127は、ワークエリア130に格納されている漢字文字列の先頭から一文字ずつずらしながら、全ての2-gramを特徴文字列として抽出する。図11は、図10の例で抽出された漢字文字列から、漢字文字列対応特徴文字列抽出プログラム127により特徴文字列を抽出した結果を示している。例えば、同一文字種文字列1000の中で文書2から抽出された「新」「しい」「ソフトウェア」「の」「開発作業」からは、「開発」「発作」「作業」が抽出される。

【0110】カタカナ文字列対応特徴文字列抽出プログラム128は、ワークエリア130に格納されているカタカナ文字列そのものを特徴文字列として抽出する。図12は、図10の例で抽出されたカタカナ文字列から、カタカナ文字列対応特徴文字列抽出プログラムにより特徴文字列を抽出した結果である。例えば、同一文字種文字列1000の中で文書2から抽出された「新」「しい」「ソフトウェア」「の」「開発作業」からは、「ソフトウェア」が抽出される。

【0111】以上が、第一の実施例における特徴文字列抽出プログラム116のステップ702における漢字文字列対応特徴文字列抽出プログラム127と、ステップ703におけるカタカナ文字列対応特徴文字列抽出プログラム128の処理手順である。

【0112】この例では、漢字文字列対応特徴文字列抽出プログラム127の処理として、漢字文字列から2-gramを特徴文字列として抽出するものとして説明したが、1-gram、あるいは3-gram以上であっても、さらには、それらの組み合わせであっても、同様に特徴文字列抽出の処理を行うことができることは明らかであろう。

【0113】次に、図4に示したステップ404でシステム制御プログラム110により起動される検索制御プログラム118による類似文書検索の処理手順について、図13のPAD図を用いて説明する。

【0114】検索制御プログラム118は、まずステップ1300で検索条件式解析プログラム119を起動し、キーボード101から入力された検索条件式を解析し、検索条件式のパラメータとして指定された種文書番号を抽出する。

【0115】次に、ステップ1301で類似文書検索プログラム120を起動し、上記ステップ1300で抽出された種文書番号に対し、磁気ディスク装置105に格納されているテキスト103中の各文書の類似度を算出する。

【0116】そして、ステップ1302において、類似度ソートプログラム126を起動し、上記ステップ1301で算出された各文書の類似度を降順にソートする。

【0117】最後に、ステップ1303において上記ステップ1302でソートされた類似度を各文書番号と共に出力する。

【0118】以上が、検索制御プログラム118による文書検索の処理手順である。

【0119】次に、図13に示したステップ1301で検索制御プログラム118により起動される類似文書検索プログラム120の処理手順について、図14のPAD図を用いて説明する。類似文書検索プログラム120は、まずステップ1400で種文書読み込みプログラム121を起動し、検索条件式解析プログラム119によって検索条件式から抽出された文書番号の種文書をワークエリア130に読み込む。ここで、種文書は、テキスト103中に格納されている文書を読み込むだけでなく、フロッピディスク107、CD-ROM装置（図1には示していない）や通信回線等を用いて、他の装置から入力するような構成を取ることも可能であり、また、全文検索システム等による検索結果から入力するような構成を取ることも可能であり、類似度ソートプログラム126の出力から種文書を選択する構成を取ることも可能である。

【0120】次に、ステップ1401において、同一文

字種文字列抽出プログラム115を起動し、上記種文書読み込みステップ1400で読み込んだ種文書のテキストを、文字種境界で分割して同一文字種文字列を取得し、ワークエリア130に格納する。

【0121】そして、ステップ1402において特徴文字列抽出プログラム116を起動し、上記同一文字種文字列抽出ステップ1401で取得した同一文字種文字列から、特徴文字列を抽出する。

【0122】図15に、この処理の具体例を示す。特徴文字列抽出プログラム116の処理手順に関しては、前に説明した通りである。

【0123】本例では、種文書である文書2「新しいソフトウェアの開発作業」から、「新」「しい」「ソフトウェア」「の」「開発作業」という5個の同一文字種文字列1500が抽出されることになる。この抽出された同一文字種文字列1500の文字種にしたがって、特徴文字列を抽出する。この結果、文書2からは「ソフトウェア」「開発」「発作」「作業」の4個の特徴文字列1501が抽出される。

【0124】次に図14のステップ1403で、出現頻度計数プログラム123を起動し、上記特徴文字列抽出ステップ1402で抽出した特徴文字列の種文書内における出現頻度を計数する。

【0125】図16に、この具体例を示す。本図は、図15に例示した種文書から抽出された特徴文字列1501の出現頻度を計数した結果を示している。すなわち、「（ソフトウェア、1）」、「（開発、1）」、「（発作、1）」、「（作業、1）」という出現頻度1600が得られている。ここで、例えば（開発、1）は、特徴文字列「開発」が「1」回出現するということを示している。

【0126】次に、図14のステップ1404で、出現頻度ファイル読み込みプログラム124を起動し、上記特徴文字列抽出ステップ1402で抽出した特徴文字列の、テキスト103中の各文書における出現頻度を出現頻度ファイル104から読み込む。

【0127】図17に、この具体例を示す。ここでは、図15の例で抽出された特徴文字列1501のテキスト103中の各文書における出現頻度を、読み込んだ出現頻度ファイルから取得した結果を示している。

【0128】この例では、種文書から抽出された「ソフトウェア」「開発」「発作」「作業」という4個の特徴文字列1501の出現頻度を、出現頻度ファイル104から得る。この結果、出現頻度1700として、例えば文書3の場合「（ソフトウェア、1）」、「（開発、1）」、「（発作、0）」、「（作業、0）」という値を得ることができる。

【0129】最後に、図14のステップ1405で、類似度算出プログラム125を起動し、上記出現頻度計数ステップ1403で計数した特徴文字列の種文書内における出現頻度と、上記出現頻度ファイル読み込みステップ

1404で読み込んだ特徴文字列のテキスト103内の各文書における出現頻度から、テキスト103中の各文書との類似度を算出する。

【0130】図18に、この具体例を示す。ここでは、図16の例で計数した種文書における出現頻度1600および図17の例で取得したテキスト103中の各文書における出現頻度1700を用いて、各文書の類似度 $S(1) \sim S(4)$ を算出した結果を示している。すなわち、次のような結果が得られる。

【0131】

$S(1)=0.077$

$S(2)=1.0$

$S(3)=0.263$

$S(4)=0.148$

本実施例では、この類似度の算出に、従来技術2に開示されている式(1)を用いるが、他の方法を用いても構わない。

【0132】以上が、類似文書検索プログラム118の処理手順である。

【0133】以上が、本発明の第一の実施例である。

【0134】なお、本実施例においては、特徴文字列抽出プログラム116は、漢字対応特徴文字列抽出プログラム127およびカタカナ文字列対応特徴文字列抽出プログラム128を含む構成としたが、英字や数字等に対応した特徴文字列抽出プログラムを含む構成としてもよいし、漢字文字列対応特徴文字列抽出プログラム127あるいはカタカナ文字列対応特徴文字列抽出プログラム128を含まない構成であってもよい。

【0135】また、本実施例においては、同一文字種文字列から特徴文字列を抽出する構成としたが、特定の文字種間を境界として前後に跨る部分文字列を特徴文字列として抽出することにより、例えば、「F1」や「ビタミンC」等の文字列を検索に用いることもでき、さらに高精度な類似文書検索を実現することが可能となる。

【0136】さらに、本実施例においては、出現頻度ファイル104を図2に示した表形式で作成されるものとしたが、この方法では、データベースが大規模になるに伴い特徴文字列の種類が増加するため、出現頻度ファイル読み込みステップ1404の処理に長大な時間を要することになる。この問題は、特徴文字列に対して検索用のインデクスを付加することにより解決できる。これにより、大規模なデータベースに対しても高速な類似文書検索を実現することが可能となる。この特徴文字列に対する検索用インデクスとしては、「特開平8-329112号公報」等が開示されているような単語インデクス方式を用いることができる。

【0137】次に、本発明の第二の実施例について図19を用いて説明する。

【0138】本発明を適用した類似文書検索システムの第二例は、種文書から抽出した特徴文字列のデータベー

ス内の各文書における出現頻度の取得に、検索漏れの無い全文検索インデクスを利用するものである。これにより、本類似文書検索システムを全文検索システムと組み合わせ実現した場合に、出現頻度ファイルをもつ必要がなくなる。

【0139】すなわち、本方法によれば、第一の実施例における出現頻度ファイル104の特徴文字列の検索に全文検索インデクスを利用することができ、大規模なデータベースに対しても高速な類似文書検索を実現することが可能となる。さらに、出現頻度ファイル104を全文検索用インデクスで代用するため、第一の実施例に比べ必要となる磁気ディスク容量を削減できることになる。

【0140】本実施例は、第一の実施例(図1)とほぼ同様の構成を取るが、類似文書検索プログラム120を構成する出現頻度ファイル読み込みプログラム124が異なる。このプログラムの代わりに、図19に示すように、特徴文字列検索プログラム1900が用いられる。

【0141】以下、本実施例における処理手順のうち、第一の実施例とは異なる類似文書検索プログラム120aの処理手順について図20を用いて説明する。

【0142】ここで、第一の実施例における類似文書検索プログラム120(図14)と異なる点は、出現頻度取得ステップ2004だけである。他の処理ステップの処理手順は、第一の実施例で説明した通りである。

【0143】出現頻度取得ステップ2004では、特徴文字列検索プログラム1900を起動し、特徴文字列抽出ステップ1402で抽出された特徴文字列を全文検索システム1901で検索することにより、テキスト103内の各文書における出現頻度を取得する。

【0144】本実施例の出現頻度取得ステップ2004で用いる特徴文字列検索プログラム1900は、検索漏れがなく、かつ、各文書における出現頻度を取得できる全文検索方式であれば、どのような方式を適用しても構わない。例えば、「特開昭64-35627号公報」

(以下、従来技術3と呼ぶ)で開示されているようなn-gramインデクス方式を用いることも可能である。

【0145】この従来技術3によるn-gramインデクス方式では、文書の登録時に、データベースへ登録する文書のテキストデータからn-gramとそのn-gramのテキスト中における出現位置を抽出し、全文検索用インデクス1903として磁気ディスク装置1902に格納しておく。検索時には指定された検索ターム中に出現するn-gramを抽出し、これらに対応するインデクスを上記磁気ディスク装置1902から読み込み、インデクス中のn-gramの出現位置を比較し、検索タームから抽出したn-gramの位置関係とインデクス中のn-gramの位置関係が等しいかどうかを判定することによって、指定された検索タームが出現する文書を高速に検索する。

【0146】この方式を用いて、特徴文字列を検索ター

ムとして全文検索システム 1901 へ入力し、該検索タームの出現文書とその位置情報を取得することにより、該特徴文字列の各文書における出現頻度を求めることが可能となる。

【0147】以下、この従来技術 3 を用いた出現頻度の算出方法を図 21 を用いて具体的に説明する。なお本図では、n-gram の n の値を 2 としている。

【0148】まず、文書の登録時にデータベースに登録するテキスト 2101 がインデクス作成部 2102 に読み込まれ、n-gram インデクス 2100 が作成される。この n-gram インデクス 2100 には、テキスト 2101 に出現する全ての 2-gram とテキスト 2101 におけるその 2-gram の出現位置が格納される。

【0149】本図に示すテキスト 2101 では、「心電」という 2-gram はテキスト 2101 (文書番号「1」) の 5 文字目、15 文字目、…に現われるので、n-gram インデクス 2100 には 2-gram 「心電」とこれに対応したかたちで出現位置 { (1, 5)、(1, 15)、…} が格納される。

【0150】検索時には、まず、検索タームが n-gram 抽出部 2103 に入力され、検索ターム中に出現する全ての n-gram とその n-gram の検索タームにおける出現位置が抽出される。次に、抽出された n-gram とこれに対応する n-gram の検索タームにおける出現位置がインデクス検索部 2104 に入力される。

【0151】インデクス検索部 2104 では、検索タームから抽出された n-gram に対応するインデクスが n-gram インデクス 2100 から読み込まれ、これらのインデクスの中から文書番号が一致し、かつ検索ターム中の位置関係と同じ位置関係を持つものが抽出され、検索結果として出力される。

【0152】検索タームとして「心電図」が入力された本図の場合、まず、n-gram 抽出部 2103 において、(n-gram 「心電」、n-gram 位置「1」) と (n-gram 「電図」、n-gram 位置「2」) が抽出される。ここで、n-gram 位置「1」は検索タームの先頭、n-gram 位置「2」はその次の文字位置を示す。

【0153】次に、インデクス検索部 2104 において、n-gram インデクス 2100 から n-gram 「心電」と「電図」に対応するインデクスが読み込まれる。これらのインデクスにおける出現位置が n-gram 位置「1」と n-gram 位置「2」のように連続するものが、すなわち隣接するものが抽出され検索結果として出力される。

【0154】本図では、n-gram 「心電」の出現位置「15」と n-gram 「電図」の出現位置「16」が隣接するため、n-gram 「心電図」が文字列として存在することが分かり、文書 1 中に検索ターム「心電図」が出現することが示される。しかし、n-gram 「心電」の出現位置「5」と n-gram 「電図」の出現位置「16」は隣接していないため、この位置には検索ターム「心電図」が出

現しないことが分かる。

【0155】本方法において、検索タームとして特徴文字列入力した場合、上記インデクス検索部 2104 から検索結果として出力される出現位置を計数することにより、該当特徴文字列の出現頻度を得ることが可能となる。

【0156】以上説明したように、本実施例によれば、出現頻度ファイルの特徴文字列検索用インデクスと出現頻度ファイルの代わりに、全文検索インデクスを利用できるため、大規模なデータベースに対しても余分なファイルを増やさずに、高速に類似文書検索を実現することが可能となる。

【0157】次に、本発明の第三の実施例について図 22 を用いて説明する。

【0158】本発明を適用した類似文書検索システムの第三例は、種文書から抽出した特徴文字列の重要度を算出し、この重要度が所定値を満たす特徴文字列に限定して、データベース内の各文書における出現頻度を取得し、これに基づいて類似度を算出するものである。

【0159】すなわち、本方法は、第一の実施例における出現頻度ファイル読み込みステップ 1404 で出現頻度の取得対象とする特徴文字列数を削減することによって、類似度算出に用いる特徴文字列数を削減し、文字数の多い種文書に対しても高速な類似文書検索を実現できるようにするものである。

【0160】本実施例は、第一の実施例(図 1) とほぼ同様の構成を取るが、類似文書検索プログラム 120 が異なり、図 22 に示すように、特徴文字列選択プログラム 2200 を有する。

【0161】以下、本実施例における処理手順のうち、第一の実施例とは異なる類似文書検索プログラム 120 b の処理手順について図 23 の PAD 図を用いて説明する。

【0162】ここで、第一の実施例における類似文書検索プログラム 120 (図 14) の処理手順と異なる点は、特徴文字列選択ステップ 2300 だけである。他の処理ステップの処理手順は、第一の実施例で説明した通りである。

【0163】特徴文字列選択ステップ 2300 では、特徴文字列選択プログラム 2200 を起動し、特徴文字列抽出ステップ 1402 (特徴文字列抽出プログラム 116) で抽出した特徴文字列の重要度を算出し、所定の値を満たす文字列を類似検索用の特徴文字列として選択する。

【0164】以下、特徴文字列選択ステップ 2300 で起動される特徴文字列選択プログラム 2200 の処理手順を図 24 の PAD 図を用いて説明する。

【0165】特徴文字列選択プログラム 2200 は、まず、ステップ 2400 において特徴文字列抽出ステップ 1402 で抽出された特徴文字列を取得し、ワークエ

リア130に格納する。

【0166】次に、ステップ2401で各特徴文字列が出現する文書数を出現頻度ファイル104から取得する。

【0167】そして、ステップ2402において、所定の重要度算出式を用いて該特徴文字列の重要度を算出する。

【0168】この結果、該重要度が所定値を満たす特徴文字列に限定し、これを類似度算出用の特徴文字列として抽出する（ステップ2403）。この重要度には、従来技術2の共通性ウェイトを用いてもよい。本実施例では、重要度の算出に以下に示す式（2）を用いる。

【0169】

【数2】

数2

$$\text{特徴文字列の重要度} = 1 + \log_2 \frac{n}{\text{NumDoc}} \quad \dots \quad 2$$

【0170】ここで、nはデータベース中の文書数、NumDocは特徴文字列のデータベースにおける出現文書数を示す。この値は、特徴文字列がデータベース中の全ての文書に出現する場合に最も小さく、特定の文書に偏って出現する場合に大きくなる。

【0171】また、特徴文字列を抽出する際に基準とする閾値としては、上限とする重要度と下限とする重要度を予め定めておいてもよいし、重要度の上位k個（kは1以上の予め定められた整数）を採るものとしてもよい。

【0172】以下、図25に示す具体例で特徴文字列選択ステップ2200の処理手順を説明する。なお本図では、図15の例で抽出した特徴文字列1501を対象とし、重要度が3.0以上である特徴文字列を選択するものとする。

【0173】まず、ステップ1404（図23）でワークエリア130に読み込んだ出現頻度ファイル104から各特徴文字列の出現文書数を取得する。この例では、文書2の特徴文字列1501の各出現文書数2500として、[ソフトウェア, 2]、[開発, 3]、[発作, 2]、[作業, 2]が得られる。ここで、[ソフトウェア, 2]は、特徴文字列「ソフトウェア」がデータベース中の「2」つの文書に出現することを表す。

【0174】次に、各特徴文字列の出現文書数2500から重要度2501を算出し、重要度が3.0以上の特徴文字列を抽出する。この結果、「ソフトウェア」という1個の特徴文字列2502が類似度算出用の特徴文字列として選択されることになる。

【0175】このように、特徴文字列の個数を4個から1個に削減することができるため、類似度算出に要する時間を大幅に削減することができる。

【0176】なお、本実施例では、出現頻度ファイル104を参照して、各特徴文字列の出現文書数を取得する構成としたが、文書登録時に各文書中の特徴文字列を計数し、各特徴文字列の出現文書数を求め、これを出現文書数ファイルとして記憶しておくことにより、さらに高速に特徴文字列を選択することも可能である。

【0177】また、本実施例では、出現頻度ファイル104を参照して、各特徴文字列の出現文書数を取得し重要度を算出する構成としたが、文書登録時に各文書にお

ける特徴文字列の重要度を算出し、これを重要度ファイルとして記憶しておくことにより、さらに高速に特徴文字列を選択することが可能となる。

【0178】さらに、本実施例では、重要度の算出に特徴文字列のデータベース中の出現文書数を用いたが、例えば、特徴文字列の文字種類や文字列長、種文書内の出現頻度あるいは出現位置等の情報のいずれか一つ、あるいは、それらを組み合わせることにより算出することも可能である。

【0179】以上説明したように、本発明によれば、分かち書きのない日本語のような文書に対して、類似文書検索を行なった場合においても、種文書から文字列を機械的に抽出することにより、どんな単語についても漏れない検索を行なうことが可能となる。また、文字種に応じて特徴文字列を抽出することにより、意味のまとまった文字列を用いて検索を行なうことができるため、高精度な類似文書検索を実現することができるようになる。さらに、抽出する文字列の種類が大幅に削減されるため、高速に類似文書を検索することができるようになる。

【0180】さらに、全文検索システムと組み合わせて用いることにより、大規模な文書データベースに対しても、高速な類似文書検索が実現可能となる。

【0181】

【発明の効果】本発明によれば、単語辞書を用いずに類似文書検索を行なった場合でも、意味のまとまった文字列を用いて検索を行なうことができるため、高精度な類似文書検索を実現することができる。また、抽出する文字列の文字種に応じて最適な長さの部分文字列（n-gram）を抽出するため、高速に類似文書を検索することができるようになる。

【図面の簡単な説明】

【図1】本発明による類似文書検索システムの第一の実施例の全体構成を示す図である。

【図2】出現頻度ファイルの構成例を示す図である。

【図3】特徴文字列抽出処理の流れを示すPAD図である。

【図4】本発明の第一の実施例におけるシステム制御プログラムの処理手順を示すPAD図である。

【図 5】本発明の第一の実施例における文書登録制御プログラムの処理手順を示すPAD図である。

【図 6】本発明の第一の実施例における出現頻度ファイル作成プログラムの処理手順を示すPAD図である。

【図 7】本発明の第一の実施例における特徴文字列抽出プログラムの処理手順を示すPAD図である。

【図 8】本発明の第一の実施例における漢字文字列対応特徴文字列抽出プログラムの処理手順を示すPAD図である。

【図 9】本発明の第一の実施例におけるカタカナ文字列対応特徴文字列抽出プログラムの処理手順を示すPAD図である。

【図 10】本発明の第一の実施例における同一文字種文字列抽出プログラムの処理例を示す図である。

【図 11】本発明の第一の実施例における漢字文字列対応特徴文字列抽出プログラムの処理例を示す図である。

【図 12】本発明の第一の実施例におけるカタカナ文字列対応特徴文字列抽出プログラムの処理例を示す図である。

【図 13】本発明の第一の実施例における検索制御プログラムの処理手順を示すPAD図である。

【図 14】本発明の第一の実施例における類似文書検索プログラムの処理手順を示すPAD図である。

【図 15】本発明の第一の実施例における特徴文字列抽出プログラムの処理例を示す図である。

【図 16】本発明の第一の実施例における出現頻度計数プログラムの処理例を示す図である。

【図 17】本発明の第一の実施例における出現頻度取得ファイル読み込みプログラムの処理例を示す図である。

【図 18】本発明の第一の実施例における類似度算出プログラムの処理例を示す図である。

【図 19】本発明の第二の実施例における検索処理系のプログラム構成を示す図である。

【図 20】本発明の第二の実施例における類似文書検索プログラムの処理手順を示すPAD図である。

【図 21】本発明の第二の実施例におけるn-gramインデックスの例を示す図である。

【図 22】本発明の第三の実施例における検索処理系の

プログラム構成を示す図である。

【図 23】本発明の第三の実施例における類似文書検索プログラムの処理手順を示すPAD図である。

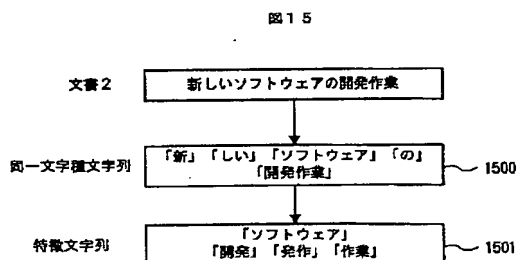
【図 24】本発明の第三の実施例における特徴文字列選択プログラムの処理手順を示すPAD図である。

【図 25】本発明の第三の実施例における特徴文字列の選択の例を示す図である。

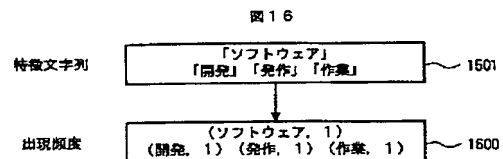
【符号の説明】

- 1 0 0…ディスプレイ、
- 1 0 1…キーボード、
- 1 0 2…中央演算処理装置（CPU）、
- 1 0 3…テキスト、
- 1 0 4…出現頻度ファイル、
- 1 0 5…磁気ディスク装置、
- 1 0 6…フロッピディスクドライブ（FDD）、
- 1 0 7…フロッピディスク、
- 1 0 8…バス、
- 1 0 9…主メモリ、
- 1 1 0…システム制御プログラム、
- 1 1 1…文書登録制御プログラム、
- 1 1 2…共有ライブラリ、
- 1 1 3…テキスト登録プログラム、
- 1 1 4…出現頻度ファイル作成登録プログラム、
- 1 1 5…同一文字種文字列抽出プログラム、
- 1 1 6…登録用特徴文字列抽出プログラム、
- 1 1 7…出現頻度ファイル作成プログラム、
- 1 1 8…検索制御プログラム、
- 1 1 9…検索条件式解析プログラム、
- 1 2 0…類似文書検索プログラム、
- 1 2 1…種文書読み込みプログラム、
- 1 2 3…出現頻度計数プログラム、
- 1 2 4…出現頻度読み込みプログラム、
- 1 2 5…類似度算出プログラム、
- 1 2 6…類似度ソートプログラム、
- 1 2 7…漢字文字列対応特徴文字列抽出プログラム、
- 1 2 8…カタカナ文字列対応特徴文字列抽出プログラム、
- 1 3 0…ワークエリア

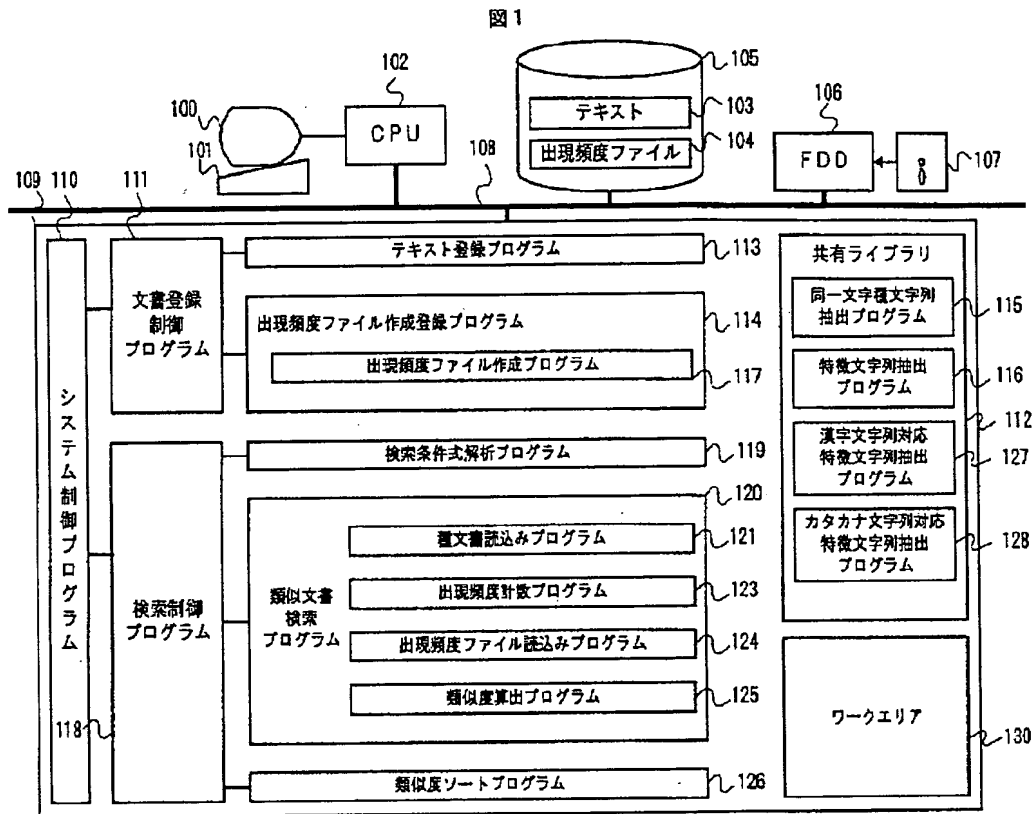
【図 15】



【図 16】



【図 1】



【図 2】

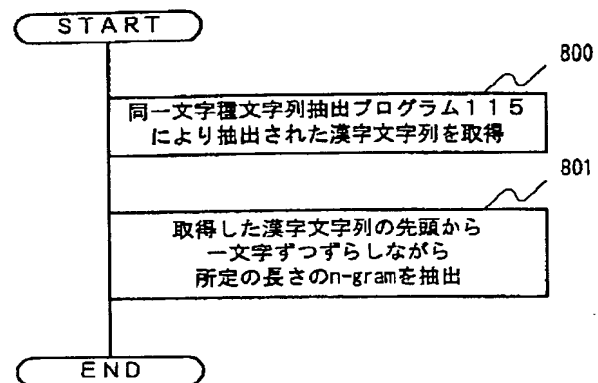
図 2

No.	特徴文字列	文書 1	文書 2	文書 3	文書 4
1	新聞	1	0	0	0
2	開発	1	1	1	0
3	心電	2	0	0	0
4	電計	1	0	0	0
5	発作	1	1	0	0
6	作時	1	0	0	0
7	電図	1	0	0	0
8	ソフトウェア	0	1	1	0
9	作家	0	1	0	1
10	ソフト	0	0	1	0
11	支援	0	0	1	0
12	ソフトクリーム	0	0	0	1
13	配布	0	0	0	1
14	布作	0	0	0	1

200

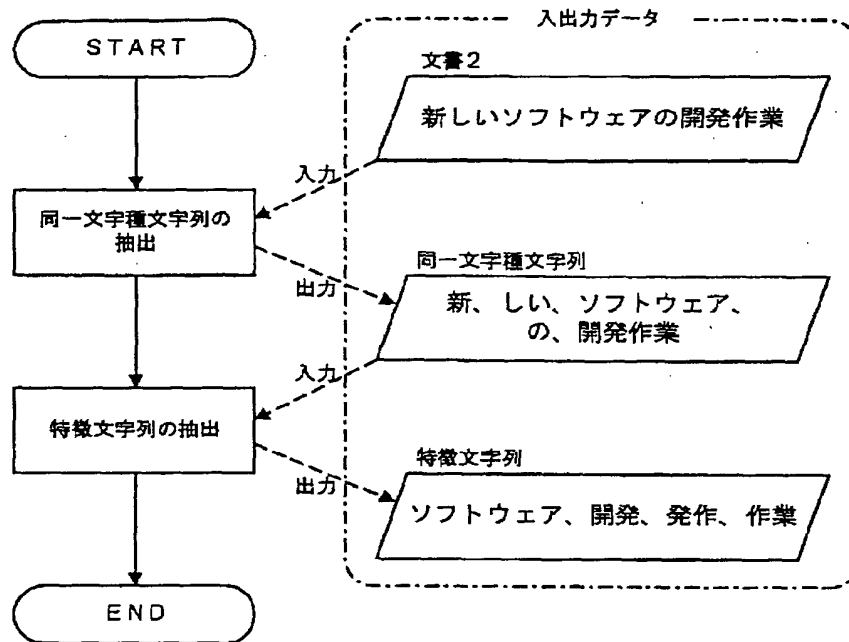
【図 8】

図 8



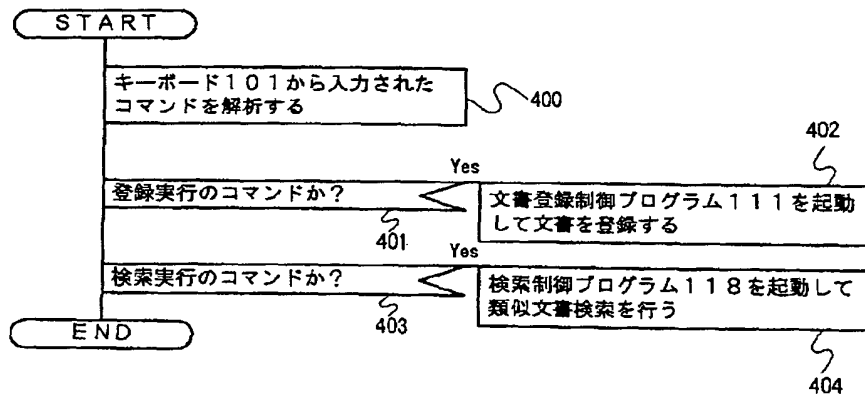
【図3】

図3



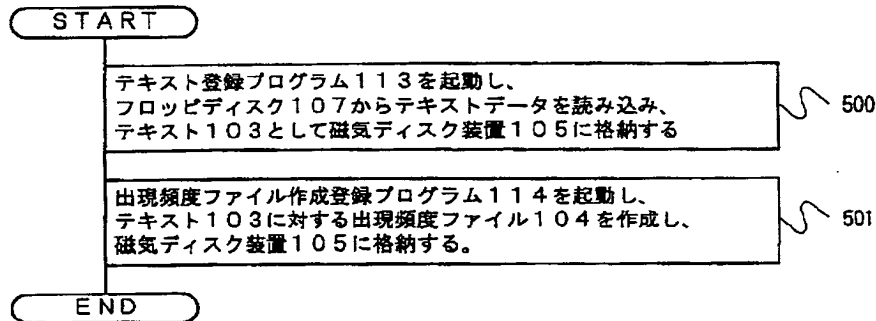
【図4】

図4



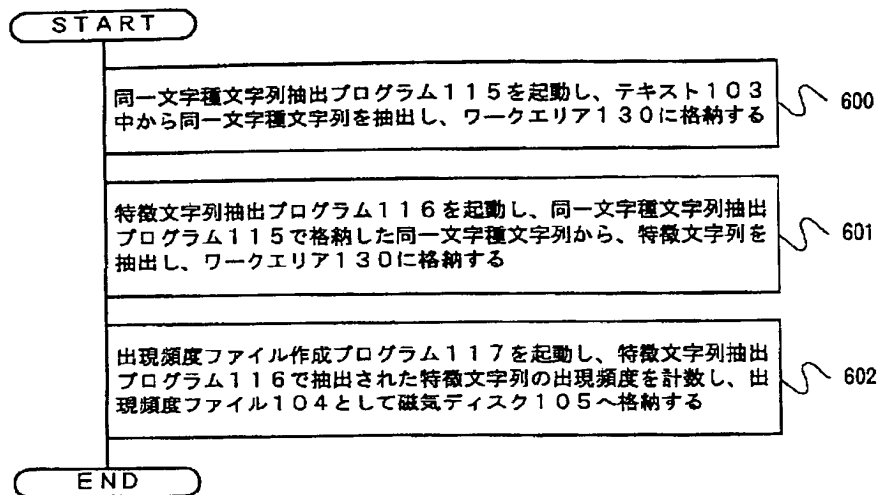
【図5】

図5



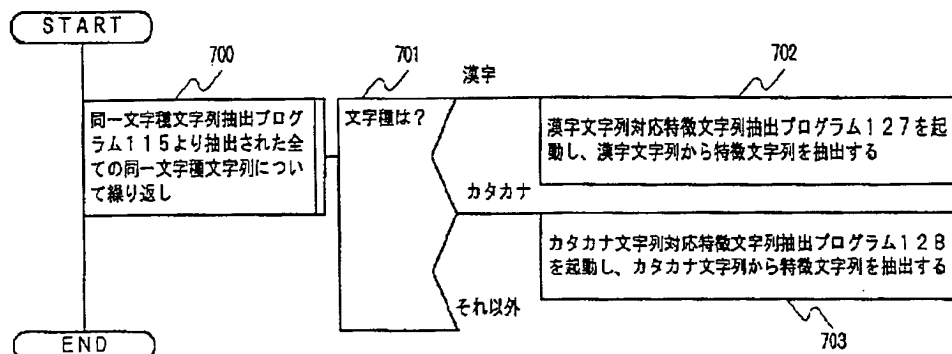
【図6】

図6



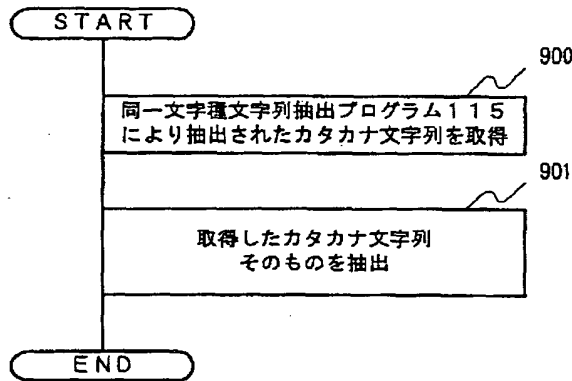
【図7】

図7



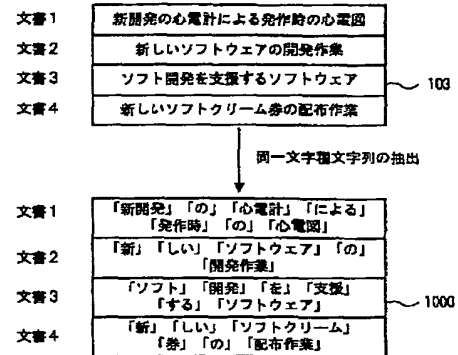
【図9】

図9



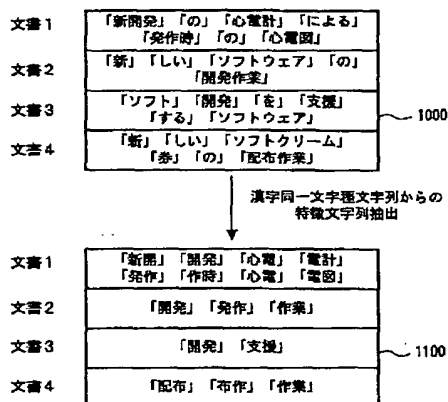
【図10】

図10



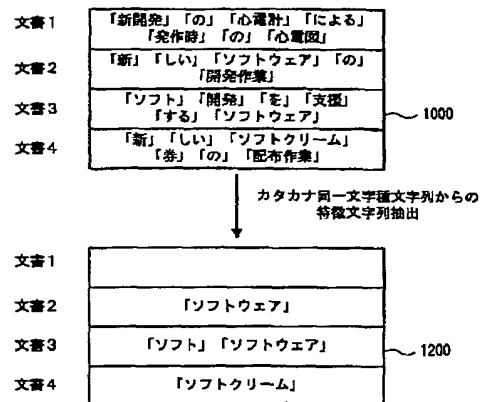
【図11】

図11



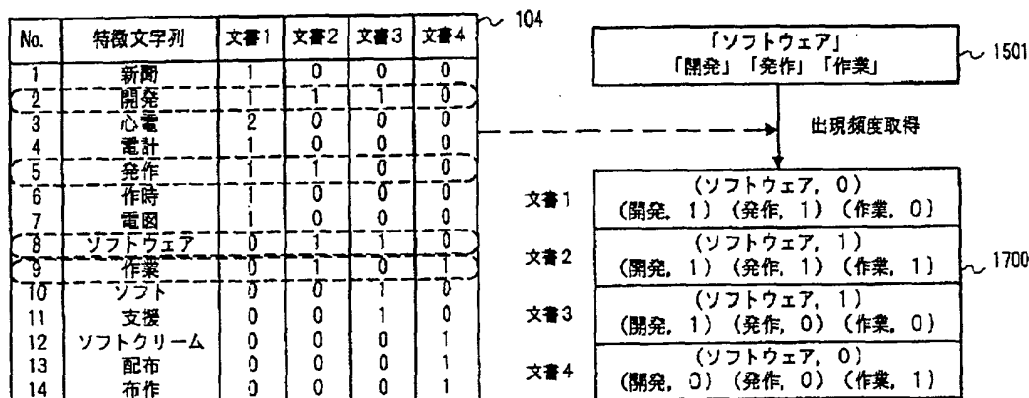
【図12】

図12

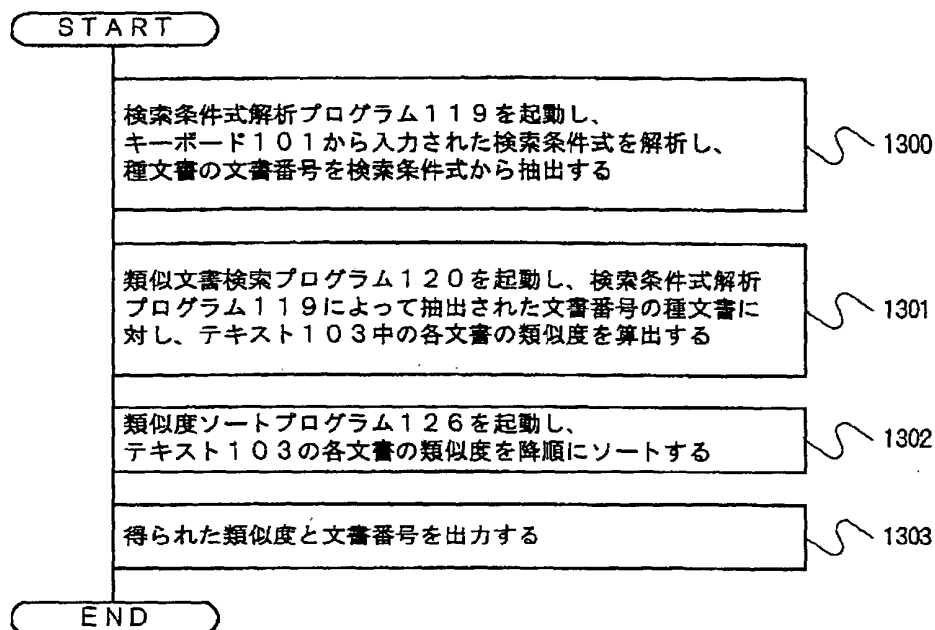


【図17】

図17



13



【图 18】

18

[illegible]

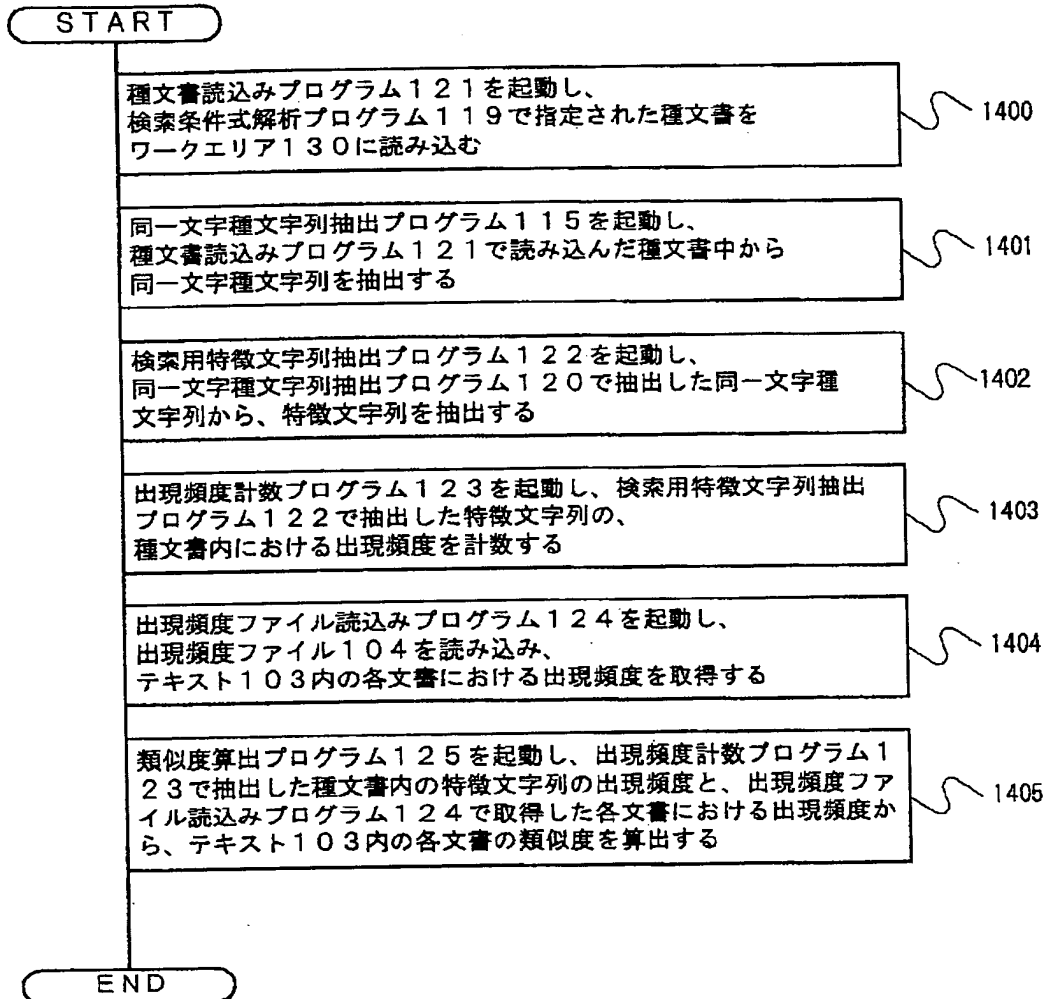
文書2の類似度	ソフトウェア		開発		発作		作業	
	文書2	種文書	文書2	種文書	文書2	種文書	文書2	種文書
出現頻度	1	1	1	1	1	1	1	1
テキスト量	14	14	14	14	14	14	14	14
ウェイト	0.071	0.071	0.071	0.071	0.071	0.071	0.071	0.071
共通性ウェイト	0.033		0.048		0.033		0.033	
類似度	1.0							

	ソフトウェア		開発		発作		作業	
	文書3	種文書	文書3	種文書	文書3	種文書	文書3	種文書
出現頻度	1	1	1	1	0	1	0	1
テキスト量	16	14	16	14	16	14	16	14
ウェイト	0.063	0.071	0.063	0.071	0	0.071	0	0.071
共通性ウェイト	0.033		0.048		0.033		0.033	
類似度	0.263							

文書4の類似度	ソフトウェア		開発		発作		作業	
	文書4	類文書	文書4	類文書	文書4	類文書	文書4	類文書
出現頻度	0	1	0	1	0	1	1	1
テキスト長	18	14	18	14	18	14	18	14
ウェイト	0	0.071	0	0.071	0	0.071	0.063	0.071
共通性ウェイト	0.033		0.048		0.033		0.033	
類似度	0.148							

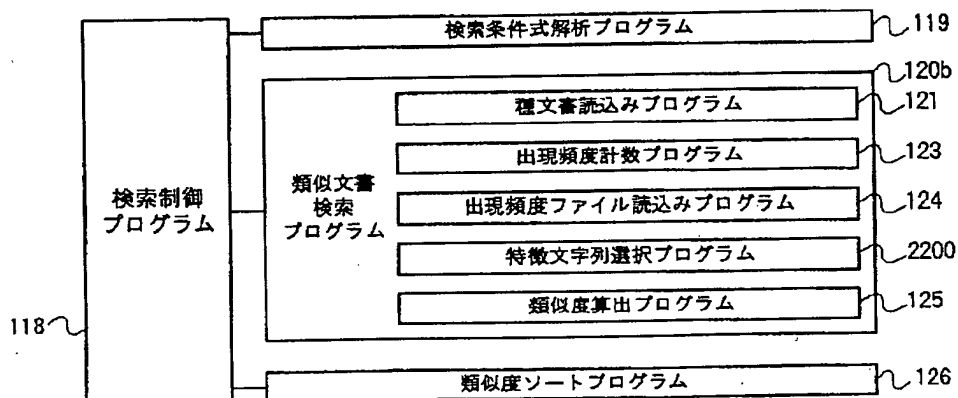
【図 14】

図 14



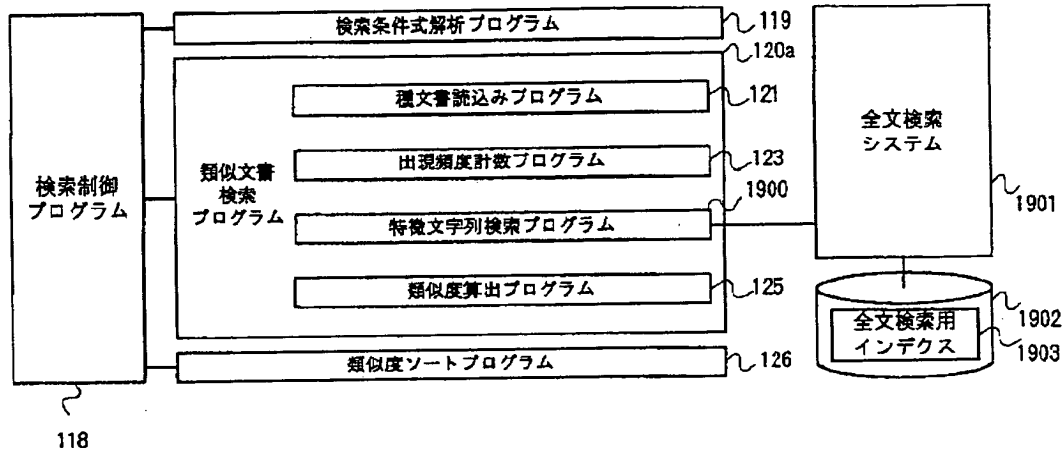
【図 22】

図 22



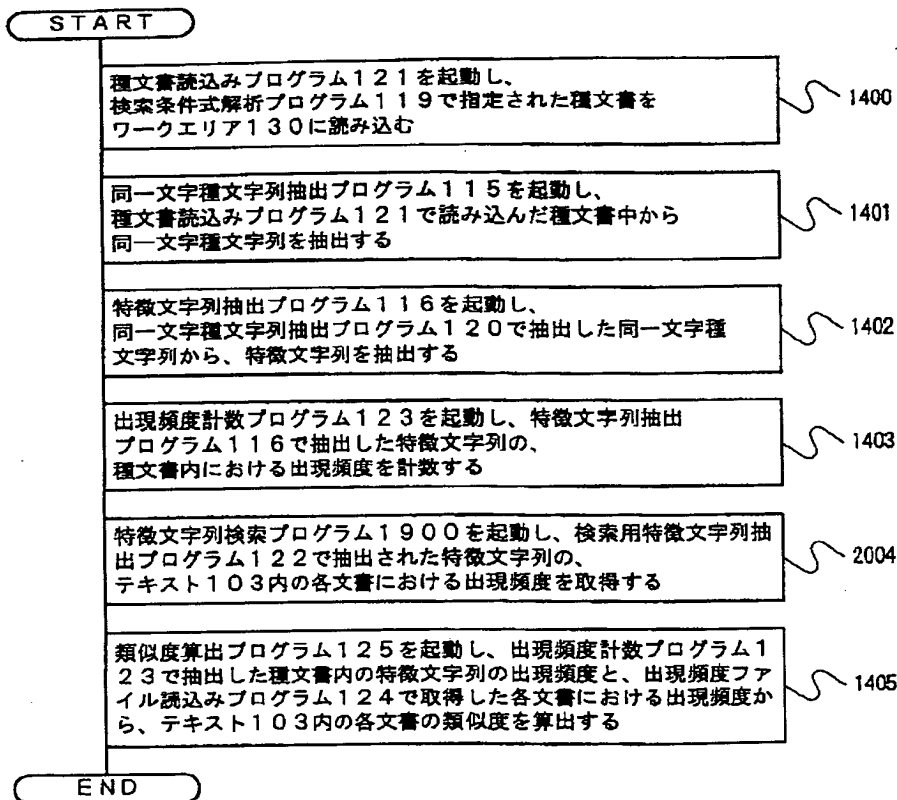
【図 19】

図 19



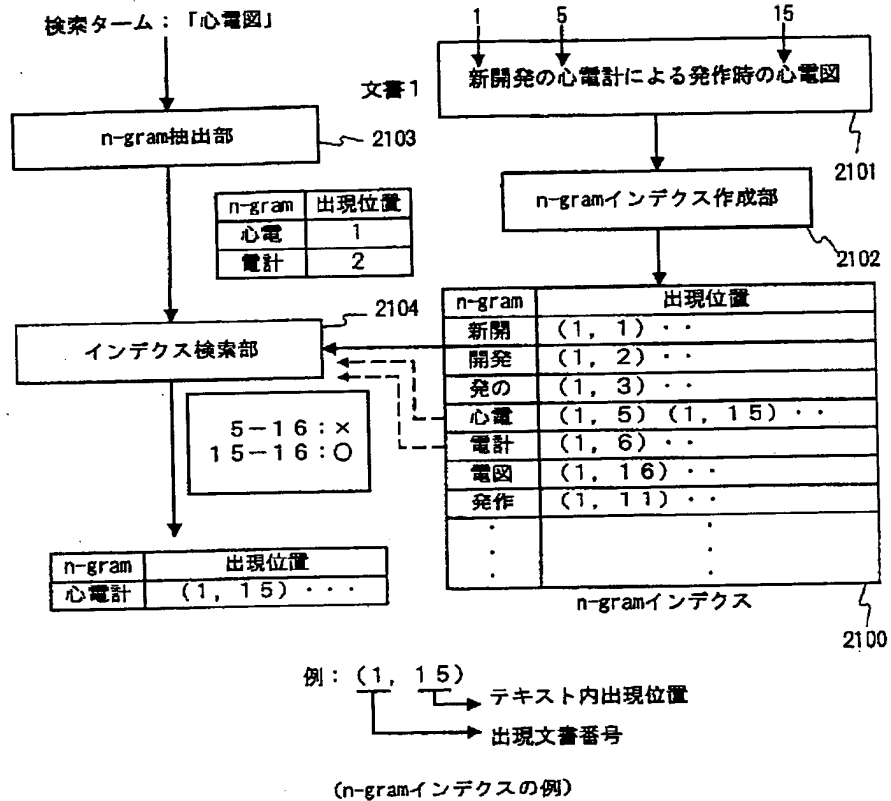
【図 20】

図 20



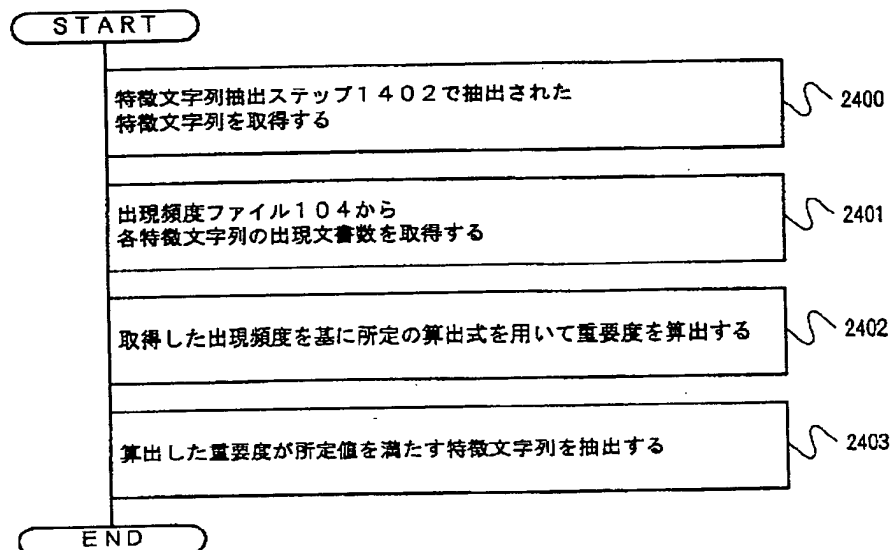
【図 2 1】

図 2 1



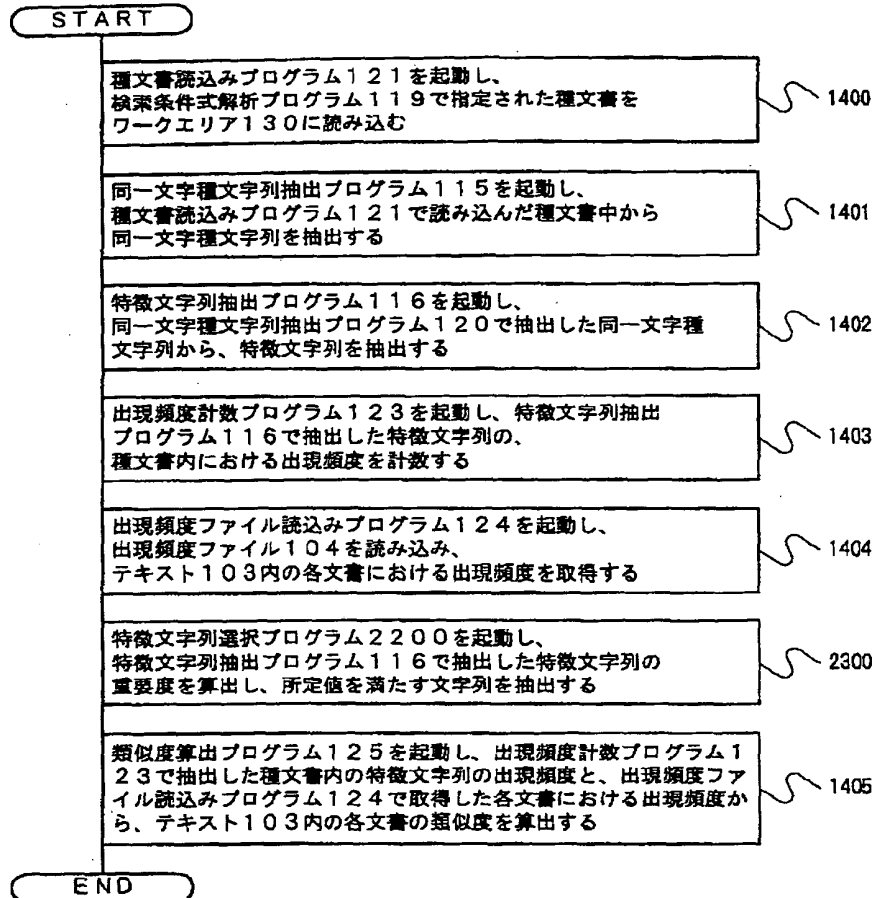
【図 2 4】

図 2 4



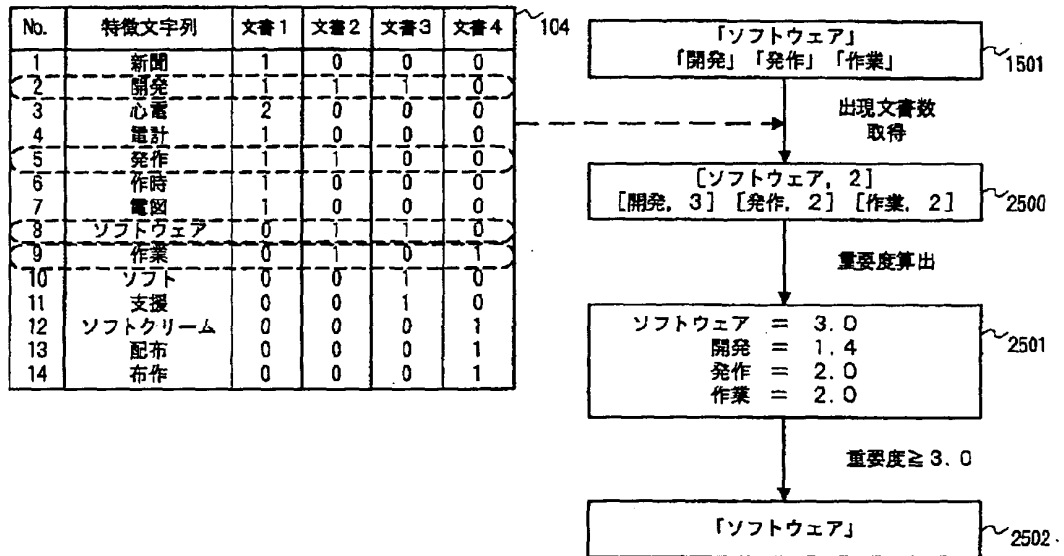
【図23】

図23



【図25】

図25



フロントページの続き

(72)発明者 菅谷 奈津子
 神奈川県川崎市幸区鹿島田890番地 株式
 会社日立製作所情報・通信開発本部内

(72)発明者 川下 靖司
 神奈川県横浜市戸塚区戸塚町3090番地 株
 式会社日立製作所ソフトウェア開発本部内

【公報種別】特許法第17条の2の規定による補正の掲載
【部門区分】第6部門第3区分
【発行日】平成14年12月20日(2002.12.20)

【公開番号】特開平11-143902
【公開日】平成11年5月28日(1999.5.28)
【年通号数】公開特許公報11-1440
【出願番号】特願平9-309078
【国際特許分類第7版】

G06F 17/30

【FI】

G06F 15/401 310 A

15/403 350 C

【手続補正書】

【提出日】平成14年9月17日(2002.9.17)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項1】文字情報をコードデータとして蓄積したテキストデータベースを対象に、ユーザが指定した文書と類似する文書を検索する類似文書検索方法において、ユーザが指定した文書のテキスト(指定テキストと呼ぶ)から所定の文字種の変わり目を境界として文字列を抽出する文字列抽出ステップと、

予め定められた一つ以上の文字列の種類に応じて、その中から一つ以上の部分文字列を抽出する検索用部分文字列抽出ステップと、

該指定テキストに対する該テキストデータベース中のテキストの類似度を所定の類似度算出式を用いて算出する類似度算出ステップを有することを特徴とした類似文書検索方法。

【請求項2】請求項1記載の類似文書検索方法における前記文字列抽出ステップは、

該指定テキストから抽出する文字列として、文字種の変わり目を境界として同一文字種からなる文字列を抽出する同一文字種文字列抽出ステップを有することを特徴とした類似文書検索方法。

【請求項3】請求項2記載の類似文書検索方法における前記検索用部分文字列抽出ステップは、

文字種に応じて予め定められた文字列長の部分文字列を検索用部分文字列として抽出する文字種別検索用部分文字列抽出ステップを有することを特徴とした類似文書検索方法。

【請求項4】請求項1、2および3に記載の類似文書検索方法における前記検索用部分文字列抽出ステップは、

予め定められた長さの文字列を検索用部分文字列として抽出する所定長文字列抽出ステップ、

前記文字列抽出ステップで抽出された文字列そのものを検索用部分文字列として抽出する最長文字列抽出ステップ、

前記文字列抽出ステップで抽出された文字列とその部分文字列の指定テキストにおける出現頻度比を算出し、所定値を満たす部分文字列を検索用部分文字列として抽出する高出現頻度比文字列抽出ステップ、

上記所定長文字列抽出ステップ、最長文字列抽出ステップおよび高出現頻度比文字列抽出ステップの中の少なくとも一つの抽出ステップで抽出された部分文字列から、予め作成しておいた、検索用部分文字列として抽出しない文字列を不要語として記載した排除文字列辞書に含まれる文字列を削除するステップ、および前記文字列抽出ステップで抽出された文字列から検索用部分文字列としては部分文字列を抽出しないステップ、

のいずれか一つ、あるいは、それらを組み合わせることにより検索用部分文字列を抽出する検索用部分文字列抽出ステップを有することを特徴とした類似文書検索方法。

【請求項5】請求項1、2、3および4に記載の類似文書検索方法はさらに、

前記検索用部分文字列抽出ステップで抽出された検索用部分文字列の重要度を、予め定められた算出式を用いて算出し、所定値を満たす検索用部分文字列を抽出する検索用部分文字列選択ステップを有することを特徴とした類似文書検索方法。

【請求項6】請求項5記載の類似文書検索方法における検索用部分文字列選択ステップは、

検索用部分文字列の抽出条件を設定する検索用部分文字列抽出条件設定ステップを有することを特徴とした類似文書検索方法。

【請求項7】請求項6記載の類似文書検索方法における検索用部分文字列抽出条件設定ステップは、

前記重要度の上限値を抽出条件として設定する重要度上限設定ステップ、
前記重要度の下限値を抽出条件として設定する重要度下限設定ステップ、
前記抽出する検索用部分文字列の個数を抽出条件として

設定する検索用部分文字列抽出個数設定ステップのいずれか 1 つ、または、これらを組み合わせることにより検索用部分文字列を抽出することを特徴とした類似文書検索方法。